

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CRISTIANE CEIA DE OLIVEIRA

SchedPingER: implementação de um workflow de tratamento de dados de desempenho de rede

RIO DE JANEIRO

2018

CRISTIANE CEIA DE OLIVEIRA

SchedPingER: implementação de um workflow de tratamento de dados de desempenho de rede

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Maria Luiza Machado Campos

RIO DE JANEIRO

2018

CIP - Catalogação na Publicação

O48s Oliveira, Cristiane Ceia de
SchedPingER: implementação de um workflow de
tratamento de dados de desempenho de rede /
Cristiane Ceia de Oliveira. -- Rio de Janeiro, 2018.
85 f.

Orientadora: Maria Luiza Machado Campos.
Trabalho de conclusão de curso (graduação) -
Universidade Federal do Rio de Janeiro, Instituto
de Matemática, Bacharel em Ciência da Computação,
2018.

1. Web Semântica. 2. Dados Abertos Conectados. 3.
Extração e transformação de dados. 4. Scheduling. I.
Campos, Maria Luiza Machado, orient. II. Título.

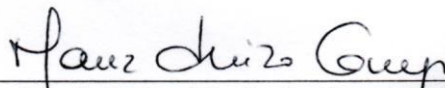
CRISTIANE CEIA DE OLIVEIRA

SchedPingER: implementação de um workflow de tratamento de dados de desempenho de rede


Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 17 de dezembro de 2018.

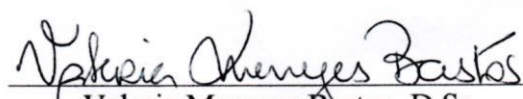
BANCA EXAMINADORA:



Maria Luiza Machado Campos, Ph.D.



Ana Carolina Brito de Almeida, D.Sc.



Valeria Menezes Bastos, D.Sc.

AGRADECIMENTOS

À minha família. Em especial, à minha mãe, pelo apoio incondicional, por não medir esforços para fazer o melhor por seus filhos e por me incentivar a sempre seguir em frente; à minha avó Helena, por toda sua dedicação, amor e cumplicidade.

Aos amigos. Por me ouvirem e oferecerem todo apoio, muitas vezes compreendendo ausências necessárias e torcendo por mim.

Aos colegas de curso que compartilharam os momentos de angústia, felicidade e estudo. Em especial, ao Armando De Luca Filho, pela amizade e pelo companheirismo desde o breve tempo como calouros da UFF.

Ao Renan Francisco Santos Souza e aos envolvidos nas pesquisas relacionadas ao GRECO e ao projeto PingER na UFRJ. A ajuda de vocês foi fundamental para a execução deste trabalho.

À minha orientadora Profa. Maria Luiza Machado Campos. Pelo conhecimento transmitido, pela confiança depositada em mim e por toda atenção e generosidade que me foram oferecidas.

A todos os meus professores, do ensino fundamental à graduação, que contribuíram para a minha formação acadêmica e pessoal. Em cada passo dessa longa jornada, há um pouco de vocês.

RESUMO

Dados de desempenho de *links* de Internet ao redor do mundo apresentam grande potencial de identificação de eventos e situações críticas quando agregados a outras bases. Este trabalho propõe uma rotina automatizada para extraí-los dos diversos arquivos em que são disponibilizados e transformá-los em triplas no formato de Dados Abertos Conectados a fim de facilitar a manipulação e a obtenção de valor a partir da correlação de dados. Para o estudo da aplicabilidade da solução, foram realizados experimentos onde se pode concluir que a conversão de todo o histórico de 20 anos de medidas de qualidade de rede é possível de forma parcial ou completa dependendo do grau de agrupamento dos dados, conforme a variação do volume de arquivos fornecidos como entrada. As principais características observadas ao elaborar uma solução como essa, bem como as dificuldades encontradas e as possibilidades de trabalhos futuros são explicitadas, evidenciando os ganhos obtidos e as possíveis melhorias do processo.

Palavras-chave: Web Semântica. Dados Abertos Conectados. Extração e Transformação de Dados. *Scheduling*.

ABSTRACT

Performance data from Internet links around the world present great potential for identifying events and critical situations when aggregated to other databases. This project proposes an automated routine to extract data from multiple files and to transform them into triples. This is done in order to facilitate both manipulation and the value obtaining through data correlations, using the Linked Open Data format. In order to analyze the applicability of the solution, experiments were carried out showing that the conversion of the entire 20-years history of network quality measurements is partially or completely possible depending on the degree of grouping of the data, according to the number of input files. The main characteristics observed when elaborating a solution such as this, as well as the difficulties faced and the possibilities of further work are explained, evidencing the achievements and the possible improvements of the process.

Keywords: *Semantic Web. Linked Open Data. Extracting and Transforming Data. Scheduling.*

LISTA DE FIGURAS

FIGURA 1 - DESCRIÇÃO DE UMA MEDIDA PINGER	20
FIGURA 2 - GRÁFICO DE QUANTIDADE DE VALORES MEDIDOS E NULOS AO LONGO DOS ANOS DE PROJETO PINGER	21
FIGURA 3 - GRÁFICO DE QUANTIDADE DE VALORES MEDIDOS POR MÉTRICA, ENTRE 1998 E 2017	22
FIGURA 4 - REPRESENTAÇÃO DE UMA TRIPLA RDF	25
FIGURA 5 - FLUXO DE TRATAMENTO DOS DADOS DO PINGER	32
FIGURA 6 - EXEMPLO DE TRIPLAS PARA UMA MEDIDA PINGER	33
FIGURA 7 - PASSO 1 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	34
FIGURA 8 - PASSO 2 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	35
FIGURA 9 - PASSO 3 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	35
FIGURA 10 - PASSO 4 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	35
FIGURA 11 - PASSO 5 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	36
FIGURA 12 - PASSO 6 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	36
FIGURA 13 - PASSO 7 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	37
FIGURA 14 - PASSO 8 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	37
FIGURA 15 - PASSO 9 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	37
FIGURA 16 - PASSO 10 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	38
FIGURA 17 - PASSO 11 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	38
FIGURA 18 - PASSO 12 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	39
FIGURA 19 - PASSO 13 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA	39
FIGURA 20 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DOS NÓS	41
FIGURA 21 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DOS CONTINENTES	41
FIGURA 22 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DOS PAÍSES	42
FIGURA 23 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DOS ESTADOS E CIDADES ..	44
FIGURA 24 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DAS MÉTRICAS	44
FIGURA 25 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DOS TAMANHOS DE PACOTE	45
FIGURA 26 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO DAS UNIDADES DE MEDIDA	46
FIGURA 27 - REPRESENTAÇÃO DO FLUXO ETL DE DESCRIÇÃO TEMPORAL	46

FIGURA 28 - CONTEÚDO DO ARQUIVO EXECUTÁVEL PARA SCHEDULING NO WINDOWS	47
FIGURA 29 - CONTEÚDO DO ARQUIVO EXECUTÁVEL PARA SCHEDULING NO CENTOS.....	47
FIGURA 30 – PROPRIEDADES DO PASSO 1 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “FILE”, INDICANDO A VARIÁVEL QUE CONTERÁ O CAMINHO PARA O ARQUIVO DE ENTRADA DA TRANSFORMAÇÃO	65
FIGURA 31 – PROPRIEDADES DO PASSO 1 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “CONTENT”, INDICANDO CARACTERÍSTICAS DO CONTEÚDO DO ARQUIVO, O SEU FORMATO E O CARACTERE SEPARADOR DE COLUNAS	66
FIGURA 32 – PROPRIEDADES DO PASSO 1 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “FIELDS”, EM QUE SÃO DEFINIDAS CADA COLUNA CONTIDA NO ARQUIVO DE ENTRADA, COMA INDICAÇÃO DO TIPO DO DADO E DO VALOR QUE REPRESENTA O VALOR NULO	67
FIGURA 33 – PROPRIEDADES DO PASSO 1 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “FIELDS”, EM CONTINUAÇÃO ÀS APRESENTADAS NA FIGURA 32	67
FIGURA 34 – PROPRIEDADES DO PASSO 1 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “ADDITIONAL OUTPUT FIELDS”, INDICANDO QUE O NOME DO ARQUIVO DE ENTRADA SERÁ DISPONIBILIZADO PARA O PASSO SEGUINTE	68
FIGURA 35 – PROPRIEDADES DO PASSO 2 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, IDENTIFICANDO CADA INFORMAÇÃO CONTIDA NO NOME DO ARQUIVO DE ENTRADA	68
FIGURA 36 – PROPRIEDADES DO PASSO 3 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, A FIM DE ARMAZENAR APENAS O MÊS A QUE OS DADOS SE REFEREM, DESCONSIDERANDO A INFORMAÇÃO QUANTO À EXTENSÃO DO ARQUIVO DE ENTRADA	69
FIGURA 37 – PROPRIEDADES DO PASSO 4 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRELACIONANDO CADA NÚMERO EM REFERÊNCIA A UM MÊS DO ANO COM A SUA ABREVIATURA CORRESPONDENTE	70
FIGURA 38 – PROPRIEDADES DO PASSO 5 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, A FIM DE PADRONIZAR CADA NOME DE MÉTRICA CONSIDERADA NO PROJETO PINGER	71
FIGURA 39 – CÓDIGO CONTIDO NO PASSO 6 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, PARA TRATAMENTO DAS VARIÁVEIS QUE ARMAZENAM AS MEDIDAS DOS DIAS 29, 30 E 31: CASO O VALOR SEJA NUMÉRICO, RETORNA O VALOR MEDIDO; SENÃO, RETORNA NULO	72

FIGURA 40 – PROPRIEDADES DO PASSO 7 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, EM QUE CADA URI É DESCRITO CONFORME A ONTOLOGIA DEFINIDA EM (SOUZA, 2013)	72
FIGURA 41 – PROPRIEDADES DO PASSO 7 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, EM CONTINUAÇÃO ÀS APRESENTADAS NA FIGURA 40	73
FIGURA 42 – PROPRIEDADES DO PASSO 7 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, EM CONTINUAÇÃO ÀS APRESENTADAS NA FIGURA 41	73
FIGURA 43 – DETALHAMENTO DA FÓRMULA QUE REPRESENTA CADA URI DA MEDIDA DE UM DIA (MEASUREMENTURI_DAYX), CONFORME VISTO NAS FIGURAS 41 E 42, ONDE X É UM DIA DE 1 A 31 REPRESENTADO POR UM NÚMERO DE DOIS DÍGITOS NO FINAL DA DEFINIÇÃO DO URI	74
FIGURA 44 – PROPRIEDADES DO PASSO 8 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, ONDE É VERIFICADA A EXISTÊNCIA DE MEDIDA PARA O DIA ANALISADO	74
FIGURA 45 – PROPRIEDADES DO PASSO 9 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE À ABA “MAPEAMENTO” DO STEP “MEASUREMENT DATA PROPERTY MAPPING”, EM QUE AS TRIPLAS INDICANDO QUE TRATA-SE DE UMA MEDIDA E A DE SEU RESPECTIVO VALOR SÃO CRIADAS	75
FIGURA 46 – PROPRIEDADES DO PASSO 9 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE À ABA “CAMPOS DE SAÍDA” DO STEP “MEASUREMENT DATA PROPERTY MAPPING”, INDICANDO COMO RESULTADO OS CAMPOS DE SUJEITO, PREDICADO E OBJETO QUE COMPÕEM A TRIPLA	75
FIGURA 47 – PROPRIEDADES DO PASSO 9 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE À ABA “MAPEAMENTO” DO STEP “MEASUREMENT OBJECT PROPERTY MAPPING”, EM QUE AS TRIPLAS DE CADA PARÂMETRO QUE DEFINE UMA MEDIDA SÃO CRIADAS	76
FIGURA 48 – PROPRIEDADES DO PASSO 9 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE À ABA “CAMPOS DE SAÍDA” DO STEP “MEASUREMENT OBJECT PROPERTY MAPPING”, INDICANDO COMO RESULTADO OS CAMPOS DE SUJEITO, PREDICADO E OBJETO QUE COMPÕEM A TRIPLA	76
FIGURA 49 – PROPRIEDADES DO PASSO 10 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, PARA DEFINIÇÃO DAS CONSTANTES NECESSÁRIAS NA CRIAÇÃO DAS TRIPLAS NO FORMATO N-TRIPLES RDF	77

FIGURA 50 – PROPRIEDADES DO PASSO 11 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE AO STEP “FORMULA: DATAPROP”, PARA A CONSTRUÇÃO DAS TRIPLAS ORIUNDAS DO STEP “MEASUREMENT DATA PROPERTY MAPPING”	77
FIGURA 51 – DETALHAMENTO DA REGRA PARA A CONSTRUÇÃO DAS TRIPLAS A PARTIR DOS CAMPOS DE SUJEITO, PREDICADO E OBJETO ORIUNDOS DO STEP “MEASUREMENT DATA PROPERTY MAPPING”	78
FIGURA 52 – PROPRIEDADES DO PASSO 11 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE AO STEP “FORMULA: OBJECTPROP”, PARA A CONSTRUÇÃO DAS TRIPLAS A PARTIR DOS CAMPOS DE SUJEITO, PREDICADO E OBJETO ORIUNDOS DO STEP “MEASUREMENT OBJECT PROPERTY MAPPING”	78
FIGURA 53 – PROPRIEDADES DO PASSO 12 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE AO STEP “DATAPROP TRIPLES”, EM QUE A TRIPLA CONSTRUÍDA NO STEP “FORMULA: DATAPROP” É SELECIONADA	79
FIGURA 54 – PROPRIEDADES DO PASSO 12 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, CORRESPONDENTE AO STEP “OBJECTPROP TRIPLES”, EM QUE A TRIPLA CONSTRUÍDA NO STEP “FORMULA: OBJECTPROP” É SELECIONADA	79
FIGURA 55 – PROPRIEDADES DO PASSO 13 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “FILE”, EM QUE SÃO DEFINIDOS O CAMINHO E O FORMATO DO ARQUIVO DE SAÍDA	80
FIGURA 56 – PROPRIEDADES DO PASSO 13 DA TRANSFORMAÇÃO PARA A GRANULARIDADE DIÁRIA, NA ABA “CONTENT”, EM QUE SÃO DEFINIDAS AS POSSÍVEIS CARACTERÍSTICAS DO CONTEÚDO DO ARQUIVO	80

LISTA DE TABELAS

TABELA 1 - RESULTADO DA EXPERIMENTAÇÃO COM DADOS DE GRANULARIDADE HORÁRIA	
.....	52
TABELA 2 - RESULTADO DA EXPERIMENTAÇÃO COM DADOS DE GRANULARIDADE DIÁRIA .	53

LISTA DE SIGLAS

CSV – *Comma-Separated Values*
ETL – *Extract, Transform and Load*
FTP – *File Transfer Protocol*
GRECO – Grupo de Engenharia do Conhecimento
HTML – *HyperText Markup Language*
HTTP – *HyperText Transfer Protocol*
IEPM – *Internet End-to-end Performance Monitoring*
LOD – *Linked Open Data*
PDI – *Pentaho Data Integration*
PingER – *Ping End-to-end Reporting*
RDF – *Resource Description Framework*
RTT – *Round Trip Time*
SLAC – *Stanford Linear Accelerator Center*
SQL – *Structured Query Language*
TCC – Trabalho de Conclusão de Curso
TCP – *Transmission Control Protocol*
UFRJ – Universidade Federal do Rio de Janeiro
URI – *Uniform Resource Identifier*
W3C – *World Wide Web Consortium*
XML – *eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO	14
1.1 APRESENTAÇÃO SLAC E PINGER	14
1.2 MOTIVAÇÃO	16
1.3 OBJETIVO	16
1.4 ESTRUTURA DO TRABALHO	17
2 CARACTERIZAÇÃO DO PROBLEMA E TRABALHOS ASSOCIADOS	18
2.1 CARACTERIZAÇÃO DOS DADOS PINGER	18
2.1.1 Quanto à origem	18
2.1.2 Quanto ao armazenamento e recuperação	20
2.1.3 Quanto ao volume	21
2.2 TRABALHOS ASSOCIADOS AO PINGER	22
3 REFERENCIAL TEÓRICO	24
3.1 DADOS ABERTOS CONECTADOS	24
3.2 PROCESSO DE TRIPLIFICAÇÃO	25
3.3 PROCESSO ETL	27
3.3.1 Script sob demanda	27
3.3.2 Ferramenta ETL	27
3.3.2.1 Pentaho Data Integration	28
3.4 ETL4LOD	30
3.5 SCHEDULING	30
4 SCHEDPINGER	32
4.1 PROCESSO ETL	32
4.1.1 ETL passo-a-passo	34
4.1.1.1 Granularidade diária	34
4.1.1.2 Granularidade horária	39
4.1.1.3 Parâmetros descritivos de medida	40
4.2 SCHEDULING	47
5 EXPERIMENTAÇÃO	50
5.1 OBJETIVO	50
5.2 CONFIGURAÇÃO DA EXPERIMENTAÇÃO	50

5.3 RESULTADOS	51
5.3.1 Granularidade horária	51
5.3.2 Granularidade diária	52
5.4 ANÁLISE DOS RESULTADOS	53
6 CONCLUSÃO.....	55
REFERÊNCIAS	57
APÊNDICE A – EXEMPLO DE ARQUIVO N-TRIPLES RDF: TRIPLAS REFERENTES ÀS MEDIDAS DE THROUGHPUT NA GRANULARIDADE HORÁRIA, PARA O DIA 01/07/2016	59
APÊNDICE B – DEMONSTRATIVO DOS PASSOS (“STEPS”) ELABORADOS PARA DADOS PINGER NA GRANULARIDADE DIÁRIA	65
ANEXO A – EXEMPLO DE ARQUIVO PINGER: MEDIDAS DE THROUGHPUT NA GRANULARIDADE HORÁRIA, PARA O DIA 01/07/2016 (THROUGHPUT-100-BY-NODE-2016-07-01.TXT).....	81

1 INTRODUÇÃO

1.1 APRESENTAÇÃO SLAC E PINGER

O *Stanford Linear Accelerator Center*¹ (SLAC) é um laboratório situado nos Estados Unidos, operado pela Universidade de Stanford. Foi inaugurado em 1962, como um grupo de 200 pessoas focadas no estudo da Física, a fim de construir e operar o maior acelerador linear de partículas. Em 2014, já contava com 1600 empregados, incluindo cientistas, engenheiros, técnicos e especialistas (SLAC, 2014, p. 44).

Atualmente, o SLAC realiza pesquisas em diferentes áreas, como: Química, Astrofísica, Ciência dos Materiais, Ciência Ambiental e Computação Científica. Quatro prêmios Nobel foram recebidos devido aos estudos realizados e mais de 700 artigos científicos são publicados a cada ano de pesquisa no laboratório (SLAC, 2016, p. 1).

Um dos grupos de pesquisa vinculados ao SLAC, o *Internet End-to-end Performance Monitoring*² (IEPM), é o idealizador do projeto *Ping End-to-end Reporting*³ (PingER). Este teve origem em 1995, com o objetivo de monitorar a qualidade de rede ao redor do mundo. Liderado por Les Cottrell – atual chefe do Departamento de Redes e Telecomunicações do SLAC – o projeto disponibiliza dados de desempenho de *links* de Internet desde 1998 até os dias atuais, obtidos por meio de *Pings*.

Ping é o mecanismo que permite medir quanto tempo é necessário para um pacote de informações ir até o seu destino e voltar ao seu ponto de origem. No projeto PingER, essa medição é feita a cada 30 minutos, ao enviar pacotes de cada um dos seus 80 nós fonte para seus mais de 700 nós destino, distribuídos em mais de 160 países.

A partir desses dados, podem ser obtidas informações úteis em diversos aspectos, como os citados por Cottrell (2011, p. 2). São eles:

- Técnico: facilitar o monitoramento de medidas de rede como tempos de perda e de resposta e taxa de transferência de dados, medindo a qualidade de um *link* específico.
- Econômico: fazer recomendações como investir no aumento de banda de Internet de um determinado lugar, baseado na análise dos dados do PingER.

¹ <<https://www.slac.stanford.edu/>>. Acesso em: 6 dez. 2018.

² <<http://www-iepm.slac.stanford.edu/>>. Acesso em: 6 dez. 2018.

³ <<http://www-iepm.slac.stanford.edu/pinger/>>. Acesso em: 6 dez. 2018.

- Detecção de problemas: o PingER pode ser utilizado para discernir se o problema é relacionado à rede, para identificar quando o problema começou, se ele continua ocorrendo, etc.
- Colaborativo: para a colaboração entre cientistas, é necessário um certo nível de qualidade de *links* de Internet. O PingER permite medir a qualidade desses *links*.
- Quantificação do impacto de eventos: o PingER tem sido usado para mostrar o impacto na performance de Internet quando ocorrem novas conexões, quando ocorrem terremotos, tsunamis, etc.
- Roteamento: o PingER pode ser usado para auxiliar na localização de roteamentos inapropriados. Também pode ser utilizado para selecionar a melhor rota.

Como exemplo de conclusões realizadas a partir da análise sobre os dados do PingER, seguem alguns casos de estudo encontrados em sua documentação, segundo Cottrell (2016):

- Em 2013, a partir das métricas do PingER, constatou-se que a Síria esteve *offline* por quase 20 horas entre os dias 7 e 8 de maio de 2013.
- Em 2012, a partir das medições referentes às taxas de transferência de dados, foi realizado um estudo no qual concluiu-se que, em 2008 – antes da Copa do Mundo e da instalação de vários cabos submarinos de fibra óptica para a África Subsaariana –, o desempenho de Internet no continente africano era equivalente ao da Europa no ano de 1994. No entanto, observada a melhoria de desempenho nos anos subsequentes, a qualidade de conexão na África conseguiria parear com a do continente europeu no ano 2028.
- Em março de 2011, ocorreu um dos piores terremotos registrados na história do Japão. Feita a análise sobre os dados do PingER, constatou-se que 1 dos 6 nós monitorados esteve inacessível por um período após o abalo sísmico. Além disto, pelos valores medidos, foi possível inferir o rompimento de cabos submarinos em que ocorrem o transporte de dados.
- Em 2006, um estudo sobre as medidas PingER relacionadas aos países da América Latina foi realizado. Através dele, observou-se que o desempenho da conexão à Internet no Brasil teve uma melhoria linear entre janeiro e julho de 2001.

1.2 MOTIVAÇÃO

Conforme apresentado na seção anterior, o PingER fornece dados que indicam a qualidade da Internet em diferentes aspectos, mundialmente. Quando agregados a outros dados, potencializam a capacidade de identificar eventos, situações ou locais críticos que precisam de um bom desempenho de rede e de acesso à informação.

Apesar desta importância e do projeto servir como contexto e ponto de partida para vários estudos, ainda não há implementada uma solução que auxilie nas etapas de extração, transformação e carga dos dados obtido a cada *Ping*.

Estes, são registrados em arquivos de extensão TXT, seguindo o padrão onde cada valor registrado está separado pelo caractere espaço. Não há um fluxo de trabalho sistematizado para processá-los e transformá-los. As aplicações analíticas são feitas *ad hoc*, conforme uma determinada necessidade.

No entanto, após duas décadas de monitoramento ponto-a-ponto ao redor do mundo, a manipulação de tal volume de dados dispostos em inúmeros arquivos torna-se um desafio. Extrair informações que auxiliem na tomada de decisões transforma-se em uma tarefa extremamente complexa.

Dada a sua relevância, o grande volume de dados multidimensionais e o potencial de agregação a outros dados, o projeto PingER torna-se objeto deste trabalho por permitir que abordagens de tratamento e processamento voltados para este contexto sejam estudadas.

Ademais, há o fato de trabalhar em um cenário real e a interação com os demais alunos, professores e pesquisadores tanto do Grupo de Engenharia do Conhecimento (GRECO), na Universidade Federal do Rio de Janeiro⁴ (UFRJ) quanto da equipe do SLAC *National Accelerator Laboratory*.

1.3 OBJETIVO

O objetivo geral deste trabalho é elaborar uma rotina automatizada de tratamento dos dados do PingER, através da construção de um fluxo de trabalho que os transforme em uma base de dados pronta para ser armazenada e publicada. Assim tratados, esses dados podem ser utilizados em estudos, pesquisas ou análises que se façam necessárias, considerando o grande potencial informacional quando estes dados são agregados a outros.

⁴ <<https://www.ufrj.br/>>. Acesso em: 6 dez. 2018.

Dessa forma, partimos dos arquivos brutos disponibilizados pela equipe do SLAC, visando transformá-los em triplas no formato de Dados Abertos Conectados (LOD – *Linked Open Data*) para que estes sejam extraídos e correlacionados de forma explícita na *Web*.

Como os dados do PingER são fornecidos em diversos modos de agrupamento, o foco do trabalho é apenas em dois deles: os que representam as medições a cada hora do dia e a cada dia do mês. Por isto, desejamos também investigar a viabilidade de realizar a transformação das medições realizadas em 20 anos de projeto PingER de forma completa ou parcial com a solução elaborada.

1.4 ESTRUTURA DO TRABALHO

O Capítulo 1 apresenta uma breve história sobre o SLAC e o PingER, objeto de estudo deste trabalho, assim como a motivação para fazê-lo e o objetivo a ser atingido.

No Capítulo 2, os dados do PingER são apresentados em detalhe e os trabalhos associados ao projeto são explicitados.

O Capítulo 3 consiste em apresentar o conceito de Dados Abertos Conectados, o processo de triplificação, o processo ETL (*Extract, Transform and Load*) e a ferramenta utilizada, bem como o conjunto de plugins ETL4LOD e a rotina de *scheduling*, necessários para o entendimento da solução implementada.

O Capítulo 4 descreve os pormenores da elaboração dos fluxos e rotinas de transformação dos dados do projeto. Ou seja, é o detalhamento da solução denominada SchedPingER.

No Capítulo 5, há a análise da viabilidade e eficácia do SchedPingER, baseada nos testes realizados a partir de amostras de dados.

Por fim, o Capítulo 6 apresenta uma visão geral do trabalho, os resultados e as dificuldades encontradas, bem como a sugestão de próximos passos relacionados ao projeto.

2 CARACTERIZAÇÃO DO PROBLEMA E TRABALHOS ASSOCIADOS

Neste capítulo são dispostas as características dos dados utilizados como insumo para a solução proposta, bem como os trabalhos realizados a partir do projeto PingER, dada a sua relevância.

2.1 CARACTERIZAÇÃO DOS DADOS PINGER

2.1.1 Quanto à origem

Os dados do PingER representam valores de medidas de qualidade de rede ao redor do mundo. Eles são gerados através de *Pings*, quando cada um dos seus 80 nós fonte enviam pacotes de dois diferentes tamanhos (100 e 1000 *bytes*) para cada um dos seus mais de 700 nós destino. Esse mecanismo é repetido em intervalos de 30 minutos e considera 16 métricas, que auxiliam na posterior análise de desempenho de rede. São elas:

- *Mean Opinion Score (MOS)*: métrica de qualidade de voz usada pela indústria de telecomunicações. Seu valor varia entre 1 e 5, onde 1 representa a qualidade mais baixa.
- *Directivity*: identifica se a conexão entre o par de nós fonte e destino apresenta pontos intermediários. É representada por um coeficiente *Alpha*. Quanto mais próximo de 1, menos indireta a rota.
- *Average Round Trip Time (Average RTT)*: é o valor médio de RTTs⁵ medidos em um período de tempo.
- *Maximum Round Trip Time (Maximum RTT)*: é o valor máximo entre os RTTs medidos em um período de tempo.
- *Minimum Round Trip Time (Minimum RTT)*: é o valor mínimo entre os RTTs medidos em um período de tempo.
- *Conditional Loss Probability*: é a probabilidade de, se um pacote enviado for perdido, perder o próximo também.

⁵ No PingER, o RTT está relacionado à distância entre os nós mais o atraso ao longo do caminho entre esses nós.

- *Duplicate Packets*: mede a quantidade de respostas duplicadas à transmissão de um pacote.
- *Inter Packet Delay Variation (IPDV)*: também conhecida como *Jitter*. Mede a variação do atraso na entrega dos pacotes. Pode indicar congestionamento ou insuficiência da largura de banda de rede para lidar com o tráfego de dados.
- *Packet Loss*: indica a percentagem de pacotes perdidos.
- *Minimum Packet Loss*: indica a menor percentagem de pacotes perdidos em um período de tempo.
- *TCP Throughput*: indica a taxa de transferência de pacotes de dados, em *kbits/s*.
- *Unreachability*: indica a proporção de períodos de tempo sem resposta de um determinado nó.
- *Zero Packet Loss Frequency*: indica a frequência em que um nó encontra-se desocupado.
- *Inter Quartile Range (IQR)*: métrica de dispersão estatística, baseada na divisão em quartis.
- *Out of Order Packets*: indica a fração de pacotes que são recebidos fora de ordem.
- *Ping Unpredictability*: é um indicador percentual da imprevisibilidade do desempenho de um *Ping*.

Mais informações sobre as métricas do PingER podem ser encontradas no *Tutorial on Internet Monitoring and PingER at SLAC*⁶.

Apesar dos *Pings* serem realizados a cada 30 minutos, o menor grão armazenado nos arquivos de dados do PingER refere-se ao dado na granularidade horária. Outras agregações temporais são feitas. Assim, o nível de detalhamento temporal pode ser observado em diversas granularidades: horária, diária, mensal, anual, últimos 60 dias, últimos 120 dias e últimos 365 dias.

Em resumo, uma medida de qualidade de desempenho de rede é o conjunto de 6 parâmetros, como mostrado na Figura 1.

⁶ <<http://www.slac.stanford.edu/comp/net/wan-mon/tutorial.html>>. Acesso em: 7 dez. 2018.

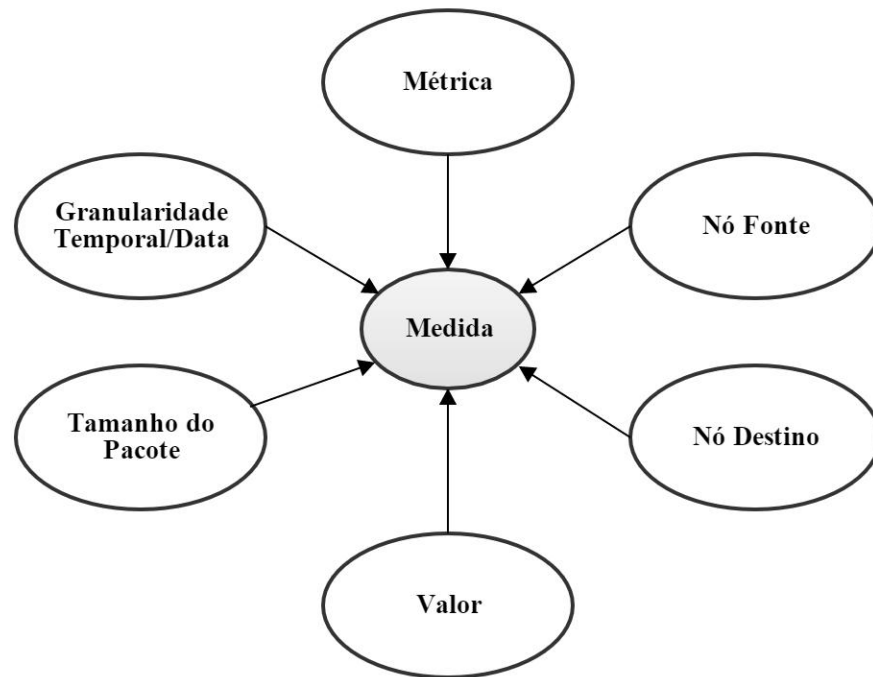


Figura 1 - Descrição de uma medida PingER

2.1.2 Quanto ao armazenamento e recuperação

Todos os dados do PingER são armazenados em múltiplos arquivos texto comum (TXT), semi-estruturados, onde cada um corresponde aos dados para uma determinada métrica, granularidade temporal e tamanho de pacote. Além dos valores medidos, contêm os nós fonte e destino considerados em cada *Ping*. No Anexo A, um exemplo do conteúdo desses arquivos pode ser visto.

Estes contêm medições a partir de 1 de janeiro de 1998 e estão disponíveis para *download* no FTP público⁷ de Les Cottrell.

Além disso, os dados podem ser consultados através da *Pingtable*⁸, aplicação que provê uma *interface Web*, onde é possível especificar os parâmetros a serem utilizados na recuperação dos dados desejados. Como resultado, é exibida uma tabela em formato HTML na página *Web*, com a possibilidade de fazer o *download* do conteúdo por meio de um *link*. No entanto, por este mecanismo, é possível a obtenção dos dados de forma pontual, limitado a um cenário específico que representa apenas uma fração do total dos dados disponíveis.

⁷ <<ftp://ftp.slab.stanford.edu/users/cottrell>>. Acesso em: 7 dez. 2018.

⁸ <<http://www-wanmon.slab.stanford.edu/cgi-wrap/pingtable.pl>>. Acesso em: 7 dez. 2018.

2.1.3 Quanto ao volume

Em julho de 2014, todos os dados referentes ao PingER – considerando os consolidados em hora, dia, mês e ano – somavam 600 *gigabytes* e eram distribuídos em aproximadamente 400 mil arquivos.

Para melhor compreensão do volume de dados do PingER, um estudo foi realizado a fim de contabilizar as medidas na granularidade horária, uma vez que esta representa a maior parcela do volume de medições, bem como o menor agrupamento de dados disponível. Foram considerados valores obtidos entre 1 janeiro de 1998 e 31 de dezembro de 2017, para todas as métricas, pares de nós e pacotes de tamanho 100 *bytes*, contidos nos 116.140 arquivos disponibilizados no FTP público de Les Cottrell.

Como resultado, obtivemos a quantidade de medidas por ano observadas no gráfico da Figura 2, totalizando mais de 8 bilhões de medidas de granularidade horária em um período de 19 anos de existência do projeto.

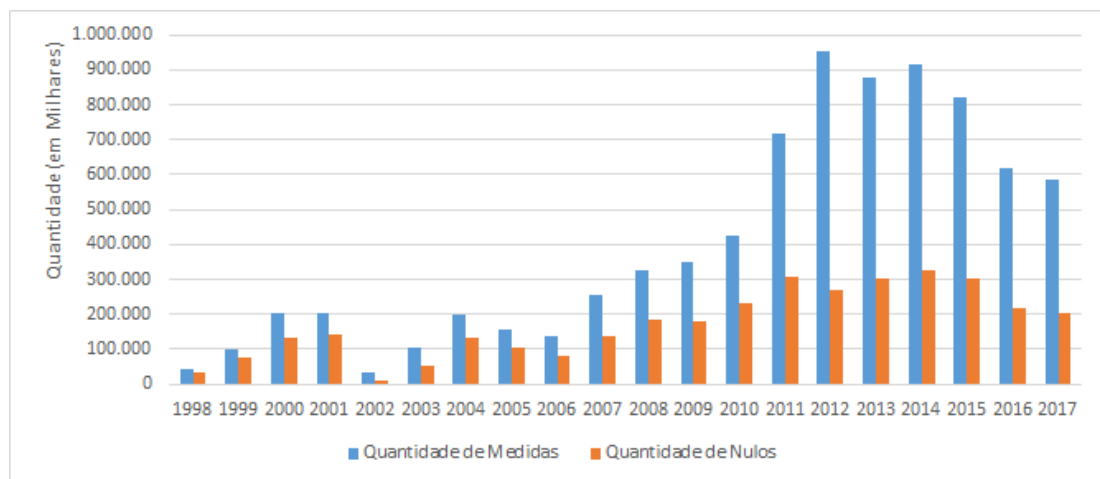


Figura 2 - Gráfico de quantidade de valores medidos e nulos ao longo dos anos de projeto PingER

Conforme é possível observar no gráfico da Figura 3, as métricas *Maximum Round Trip Time (Maximum RTT)*, *Throughput* e *Directivity* (representada pelo coeficiente *Alpha*), respectivamente, são as que mais contribuem para o volume de valores medidos.

Cabe ressaltar que os números resultantes do estudo realizado representam apenas uma fração da base de dados. Caso fossem consideradas outras granularidades (diária, mensal, anual, etc), esse volume seria ainda maior.

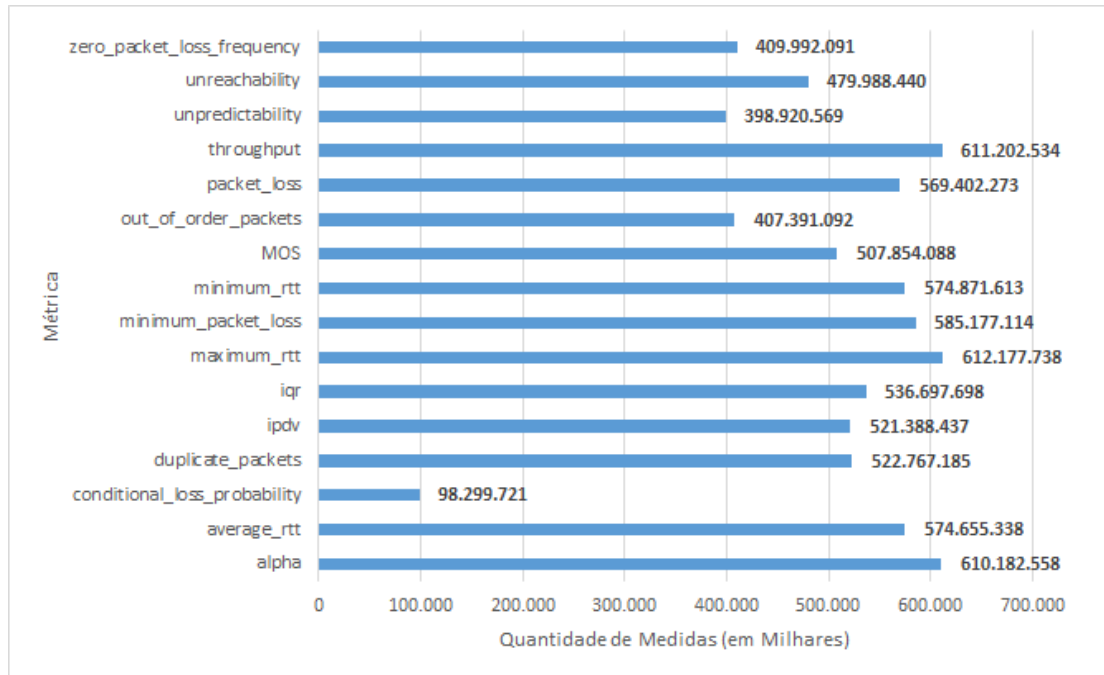


Figura 3 - Gráfico de quantidade de valores medidos por métrica, entre 1998 e 2017

2.2 TRABALHOS ASSOCIADOS AO PINGER

O PingER é usado como base de estudos por cientistas e pesquisadores de diversos países, como: Estados Unidos, Paquistão, Malásia e Brasil. No Brasil, o estudo iniciou-se através de uma parceria entre o GRECO/UFRJ, representado pela Profa. Maria Luiza Machado Campos, e o SLAC, operado pela Universidade de Stanford, representado pelo Prof. Les Cottrell.

Em 2013, Renan Francisco Santos Souza, então aluno de graduação em Ciência da Computação na UFRJ, trabalhou como estagiário em pesquisa no SLAC, junto ao projeto PingER. Sua pesquisa propôs a publicação dos dados do PingER como Dados Abertos Conectados (*Linked Open Data* – LOD), visando provê-los de forma padronizada e facilitar a interoperabilidade de bases, possibilitando cruzar dados de natureza e fontes diferentes. Assim, surgiu o PingER LOD (*PingER Linked Open Data*), o qual transformou-se em Trabalho de Conclusão de Curso (TCC). Como resultado, um *website* foi construído para descrições e informações sobre o projeto, consultas e visualizações dos dados do PingER em LOD (SOUZA, 2013).

Em 2014, surgiu o MultiLod, grupo de pesquisa ligado ao GRECO, com o objetivo de experimentar diversas abordagens de tratamento e manipulação de um grande volume de dados. Tendo como base o PingER, dividiu-se em linhas de pesquisa diferentes, as quais

originaram trabalhos diversos defendidos ao final de graduações e mestrados. Este trabalho trata-se de um deles.

Uma das linhas de pesquisa foi conduzida pela Raphaela Pedreira Nunes, em sua dissertação de mestrado em Sistemas de Informação pela UFRJ. Seu objetivo foi desenvolver um ambiente para apoiar a execução de consultas analíticas distribuídas aos dados do PingER em LOD. Para isso, foi implementado um tradutor que transforma consultas SPARQL⁹ em consultas SQL¹⁰, uma vez que o processamento distribuído ocorre em um *cluster* Hadoop¹¹, utilizando como meio de armazenamento o Impala¹² (sistema de gerenciamento de banco de dados analítico). Além disso, uma nova forma de descrever as medidas PingER em LOD foi estudada. Esta alternativa visa melhorar o desempenho de consultas sobre o volume de dados do PingER (NUNES, 2016).

Outra linha considerada, foi conduzida por Bernardo Saab Martiniano de Azevedo, Bacharel em Ciência da Computação pela UFRJ. Ele analisou duas formas de representação de dados, utilizando a base do PingER para experimentação. Em seu TCC, comparou o desempenho entre dois bancos de dados: o MySQL¹³, representando os dados em modelo relacional; e o Apache Cassandra¹⁴, para os mesmos dados em modelo não-relacional. A partir desta análise, investigou solução adequada para uma aplicação de publicação de dados, a fim de estimular futuras implementações que integrem mais fontes de dados (AZEVEDO, 2016).

Por fim, em 2015, o aluno Thiago Barbosa, da primeira turma de graduação em Sistemas de Informação na Universidade Federal Rural do Rio de Janeiro (UFRRJ)¹⁵, esteve envolvido na produção do artigo científico apresentado na conferência NETAPPS2015¹⁶ que ocorreu na Malásia. Nele são apresentadas técnicas para transformar os dados do PingER em dados estruturados, armazenados em um *Data Warehouse* capaz de realizar, em segundos, consultas analíticas complexas sobre o grande volume de dados do PingER (BARBOSA, 2015).

⁹ <<https://www.w3.org/TR/rdf-sparql-query/>>. Acesso em: 7 dez. 2018.

¹⁰ <<http://www.w3schools.com/sql/>>. Acesso em: 7 dez. 2018.

¹¹ <<http://hadoop.apache.org/>>. Acesso em: 7 dez. 2018.

¹² <<http://impala.io/>>. Acesso em: 7 dez. 2018.

¹³ <<https://www.mysql.com/>>. Acesso em: 7 dez. 2018.

¹⁴ <<http://cassandra.apache.org/>>. Acesso em: 7 dez. 2018.

¹⁵ <<http://portal.ufrj.br/>>. Acesso em: 7 dez. 2018.

¹⁶ <<http://www.internetworks.my/>>. Acesso em: 7 dez. 2018.

3 REFERENCIAL TEÓRICO

3.1 DADOS ABERTOS CONECTADOS

Linked Data refere-se a um conjunto de práticas proposto por Tim Berners-Lee a fim de publicar e interligar dados de diversas fontes na *Web*, alterando significativamente a forma como é publicado e consumido o conhecimento.

Na abordagem tradicional, dados heterogêneos são disponibilizados na *Web* em formatos como CSV, XML ou páginas HTML, com pouca preocupação com sua integração ou estrutura semântica. Neste caso, a relação entre dois documentos ocorre implicitamente (BIZER; HEATH; BERNERS-LEE, 2009).

Na *Web Semântica*, a relação entre dados de fontes diversas ocorre explicitamente, uma vez que são publicados de forma a serem compreendidos também por máquinas, com ligações aos dados externos bem definidas.

Assim, em julho de 2006, Tim Berners-Lee sugeriu 4 princípios básicos para que cada dado publicado seja parte de um único e global repositório, a *Web de Dados* (BERNERS-LEE, 2006). São eles:

- Usar Identificadores Uniformes de Recurso (URIs) para nomear as coisas.
- Usar URIs HTTP para que as pessoas possam consultar o que desejam.
- Quando um URI for consultado, fornecer informações úteis, utilizando padrões.
- Incluir *links* para outros URIs, assim é possível descobrir mais informações.

O URI é uma cadeia de caracteres que identifica um recurso em toda a *Web*, portanto, deve ser único. Este, quando especificado com o esquema *http://*, além de denominar um recurso, pode indicar um endereço para localizar documentos ou outros itens relacionados.

A fim de padronizar a disponibilização desses dados, o *Linked Data* baseia-se no RDF (*Resource Description Framework*), modelo de dados baseado em grafo, para estruturar e correlacionar dados que descrevem o que se queira definir. Essa descrição ocorre através de triplas formadas por sujeito, predicado e objeto (ou recurso, propriedade e valor).

Na Figura 4, é representada a definição de tripla RDF. O componente sujeito é um URI que identifica um recurso; o objeto é um URI que identifica um recurso ou um texto

literal; por fim, o predicado é um URI que especifica como o sujeito e o objeto estão relacionados.

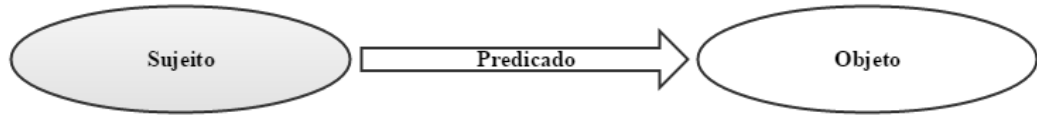


Figura 4 - Representação de uma tripla RDF

Como maior exemplo de aplicação dos princípios de *Linked Data*, há o projeto *Linking Open Data*¹⁷, iniciado em janeiro de 2007, que tem como objetivo identificar bases de dados sob licença aberta e convertê-las e publicá-las em RDF, a fim de interligar dados de natureza e fontes diferentes. Essas relações são representadas em uma nuvem LOD, ou seja, como um grande grafo que explicita as ligações entre as bases. Este diagrama, bem como cada detalhe das bases de dados representadas nele, pode ser acessado interativamente na página *The Linking Open Data cloud diagram*¹⁸.

Alguns dos conjuntos de dados com mais relações são o *DBpedia*¹⁹ e o *GeoNames*²⁰. O primeiro contempla dados estruturados extraídos de *infoboxes* localizadas no lado direito de artigos da *Wikipedia*²¹. O segundo oferece descrições estruturadas de locais pelo mundo, somando quase 3 milhões de lugares povoados e mais de 10 milhões de nomes geográficos.

3.2 PROCESSO DE TRIPLIFICAÇÃO

O processo de transformação de dados para o formato de triplas inclui cinco etapas, conforme descrito em (EQUIPE GT, 2011), sendo:

- Identificação e seleção de dados.
- Limpeza, anotação e transformação.
- Mapeamento.
- Interligação.

¹⁷ <<https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>>. Acesso em: 7 dez. 2018.

¹⁸ <<http://lod-cloud.net/>>. Acesso em: 7 dez. 2018.

¹⁹ <<http://wiki.dbpedia.org/>>. Acesso em: 7 dez. 2018.

²⁰ <<http://www.geonames.org/>>. Acesso em: 7 dez. 2018.

²¹ <<https://www.wikipedia.org/>>. Acesso em: 7 dez. 2018.

- Armazenamento e publicação.

Na fase de identificação e seleção de dados é importante considerar o tema a ser abordado e quais perguntas pretende-se responder através deles. Geralmente, as bases são disponibilizadas de forma legível à compreensão humana, porém exigindo um tratamento específico para que estas sejam compreendidas também por máquinas.

Para tal, há a etapa seguinte, de limpeza, anotação e transformação dos dados. Esta visa garantir que os dados estejam padronizados, apresentando uma estrutura que possibilite a construção automatizada de triplas no formato sujeito, predicado e objeto de forma consistente, sem que haja interpretações dúbias. Neste passo também podem ser aplicados tratamentos quanto aos caracteres especiais, acentuação e espaçamento entre palavras que compõem um dado.

A fase de mapeamento consiste em realizar a modelagem dos dados. Ou seja, a partir do entendimento das informações contidas na base de dados, definir como associá-las por meio dos predicados e quais triplas gerar. Para isto, necessita-se da definição das ontologias a serem utilizadas ou construídas.

Ontologia, conforme definido por Berners-Lee, Hendler e Lassila (2001),

É um documento ou arquivo que define formalmente as relações entre os termos. O tipo mais comum de ontologia para a *Web* tem uma taxonomia e um conjunto de regras de inferência. A taxonomia define classes de objetos e as relações entre eles.

Ou seja, ontologia é o mecanismo que permite expressar com clareza sobre qual tipo de entidade ou objeto a que estamos nos referindo, seja pelo entendimento humano ou automatizado.

Após a seleção dos dados, seu tratamento e mapeamento, chegamos à etapa de interligação. Nesta fase é aplicada a modelagem prevista anteriormente, formando as triplas que compõe a base de dados esperada em RDF. Os dados devem associar-se a um URI, a fim de garantir que estes possam ser conectados a outros disponíveis na *Web*.

Por fim, há a etapa de armazenamento e publicação, que consiste em disponibilizar os dados convertidos para que eles sejam utilizados e integrados às outras fontes. Além disso, no caso de elaboração de novos URIs, é desejável que estes sejam hospedados em páginas *Web*, contendo informações sobre o elemento.

3.3 PROCESSO ETL

ETL (*Extract, Transform and Load*) refere-se ao processo de extração, transformação e carga de dados. Essas etapas consistem em extrair os dados de uma ou mais fontes e realizar toda a transformação necessária (padronizar, tratar desvios e inconsistências, garantir a qualidade dos dados) para que eles sejam carregados em um determinado sistema de gerenciamento.

Como destacado por Kimball e Caserta (2004), este processo deve agregar valor significativo aos dados: remover incorreções e corrigir dados faltantes; prover medidas de confiança nos dados; ajustar dados de múltiplas fontes para usá-los juntos; estruturá-los para serem aproveitados em ferramentas do usuário final.

O processo ETL pode ser aplicado a um conjunto de dados através da construção de seu próprio *script*, planejando e implementando cada etapa. Ou, é possível fazê-lo ao utilizar uma das ferramentas disponíveis no mercado, voltadas para esse tipo de solução. Nas subseções 3.3.1 e 3.3.2 são apresentados prós e contras de cada uma dessas alternativas.

3.3.1 Script sob demanda

O ETL de um conjunto de dados pode ser feito através de código de programação desenvolvido especificamente para esta tarefa. Essa opção provê maior flexibilidade, uma vez que você estará programando seu próprio processo com todas as funcionalidades desejadas. Porém, implementar um sistema ETL pode não ser trivial. Além de dominar técnicas e linguagem de programação, é necessário planejar cada detalhe da execução.

Por ser uma solução voltada para um tratamento específico, admite pouco reuso. Realizar alterações no fluxo de tratamento dos dados torna-se uma tarefa trabalhosa e pouco intuitiva, dificultando a manutenção e documentação de todo o processo aplicado.

3.3.2 Ferramenta ETL

Devido à importância e à complexidade do processo ETL, estão disponíveis ferramentas de integração de dados que permitem planejar todo o processo que será aplicado a um conjunto de dados. Estas apresentam o tratamento dos dados em um fluxo visual, de forma estruturada e lógica, facilitando a compreensão do que é feito a cada etapa.

Como desvantagem nesta alternativa, pode-se imaginar que ao adotar uma ferramenta ETL perde-se a flexibilidade que um *script* sob demanda pode oferecer, uma vez que estaremos limitados à solução do fabricante do *software*. Porém, hoje esses produtos oferecem uma infinidade de recursos, além de alguns também permitirem integrar um *script* ao seu fluxo de transformação, caso ainda seja necessário.

A adoção de uma ferramenta ETL apresenta um custo financeiro caso a opção escolhida exija o pagamento de uma licença de uso. Entretanto, há disponíveis opções gratuitas, como: *Pentaho Data Integration*²², *Talend Open Studio*²³, *Jaspersoft ETL*²⁴, *KETL*²⁵, etc. Neste trabalho, optou-se pelo uso da ferramenta *Pentaho Data Integration*, que será apresentada no item 3.3.2.1, pois ela apresenta um processo contínuo de melhorias e de disponibilização de novas funcionalidades, além de permitir a inclusão de passos específicos que auxiliam no processo de triplificação de dados.

3.3.2.1 Pentaho Data Integration

Pentaho Data Integration (PDI – também conhecido como *Kettle*) é uma plataforma de integração de dados desenvolvida em Java²⁶, contida no conjunto de produtos do *Pentaho*²⁷, grupo que oferece soluções também em mineração de dados, processamento analítico, visualização de dados e *Big Data*.

Idealizado em 2003 por Matt Casters, Chefe de Integração de Dados do Grupo *Pentaho* até fevereiro de 2018, surgiu da necessidade de encontrar um modo melhor, mais rápido e menos custoso para realizar o processo ETL em um conjunto de dados. Em dezembro de 2005 tornou-se ferramenta de código aberto, isto é, o código-fonte do PDI está disponibilizado e licenciado de forma a possibilitar que este seja utilizado, modificado e distribuído gratuitamente.

O PDI possui interface gráfica baseada no modelo *drag-and-drop* que simplifica a criação de fluxos de processos ETL, além de bibliotecas de componentes padrão para acessar, preparar e combinar dados das mais diversas fontes, sejam elas arquivos ou bancos de dados.

²² <<http://www.pentaho.com/product/data-integration>>. Acesso em: 8 dez. 2018.

²³ <<https://www.talend.com/products/talend-open-studio>>. Acesso em: 8 dez. 2018.

²⁴ <<https://community.jaspersoft.com/project/jaspersoft-etl>>. Acesso em: 8 dez. 2018.

²⁵ <<http://www.ketl.org/>>. Acesso em: 8 dez. 2018.

²⁶ <<http://www.oracle.com/br/java/>>. Acesso em: 8 dez. 2018.

²⁷ <<http://www.pentaho.com/>>. Acesso em: 8 dez. 2018.

Durante o preparo da transformação, permite pré-visualizar o resultado que será obtido em cada etapa. Durante a execução, apresenta notificações e alertas sobre cada ação realizada. Assim, o entendimento sobre o que ocorre no fluxo implementado é facilitado.

As informações sobre a ferramenta e a resolução de possíveis dúvidas são acessíveis através da vasta documentação, que inclui guias explicativos e tutoriais. Além disso, há uma comunidade ativa formada por desenvolvedores, pesquisadores, especialistas e usuários do PDI que colaboram através de: discussão em fóruns, contribuições no código e na documentação, reporte de *bugs* e submissão de resultados de testes. Cabe ressaltar que funcionalidades extras podem ser incluídas no PDI através da criação de *plugins* e, estes, podem ser submetidos ao *Pentaho Marketplace*²⁸ a fim de tornar público e disponibilizar para uso o trabalho realizado.

Por fim, é necessário estabelecer alguns conceitos relacionados à ferramenta e sua arquitetura:

- Transformação: conjunto de passos interligados, formando um fluxo a fim de extrair, transformar e carregar dados. Quando armazenada em arquivo, possui extensão *.ktr*.
- *Job*: pode ser visto como coleção de transformações. É orientado a tarefas. Estas são executadas em ordem e podem ser relacionadas à implementação do ETL, nível de segurança, entre outros. Quando armazenado em arquivo, possui extensão *.kjb*.
- O PDI é composto por 4 componentes: *Spoon*, *Pan*, *Kitchen* e *Carte*.
 - *Spoon*: aplicação com interface gráfica baseada no modelo *drag-and-drop*, que permite criar e editar transformações e *jobs*. Apresenta perspectivas para executar e verificar erros em transformações e *jobs*, além de visualização e geração de modelo de dados.
 - *Pan*: aplicação que executa transformações através de linhas de comando.
 - *Kitchen*: aplicação que executa *jobs* através linhas de comando.
 - *Carte*: servidor *Web* que permite executar e monitorar remotamente transformações e *jobs*.

²⁸ <<http://www.pentaho.com/marketplace/>>. Acesso em: 7 dez. 2018.

3.4 ETL4LOD

ETL4LOD é um conjunto de *plugins* desenvolvido por pesquisadores do GRECO/UFRJ, em linguagem de programação Java, a fim de auxiliar no processo de transformação de dados em LOD através da inclusão de funcionalidades extras no PDI.

Elaborado em 2011 para ser aplicado à versão 4.1.0 *stable* da ferramenta supracitada, também conhecida como *Kettle*, tem seu código fonte aberto e disponível na plataforma *GitHub*²⁹ e possui as seguintes funcionalidades representadas em forma de *steps*:

- *Data Property Mapping*: transforma as linhas do fluxo de entrada em triplas RDF (com sujeito, objeto e predicado) para o fluxo de saída, mapeando cada um dos componentes da tripla conforme a ontologia definida pelo usuário. Neste *step*, o objeto deve ser um valor literal.
- *Object Property Mapping*: análogo ao *Data Property Mapping*, porém, o objeto deve ser um URI de um recurso.
- *Sparql Endpoint*: permite a extração de dados em formato RDF oriundos de um SPARQL *Endpoint*, conforme a definição de sua URL e a configuração da consulta pelo usuário.
- *Sparql Update Insert*: insere triplas em um repositório, através da URL do SPARQL *Endpoint* desejado.
- *Any23 Converter*: permite a conversão de dados para os formatos viáveis para LOD, são eles N-Triples, Turtle e RDF/XML.

3.5 SCHEDULING

Scheduling é o agendamento de tarefas a serem realizadas sob critério definido a partir das necessidades de cada solução. Permite controlar e otimizar a execução de processos e fluxos de trabalho de forma sistematizada, eliminando a necessidade de um *start* manual.

Ao planejar o *scheduling* de um processo ETL, deve-se avaliar qual é o melhor critério para ser o ponto de partida do agendamento. É necessário ter uma estratégia de execução bem definida, considerando as relações e dependências entre as ações do fluxo considerado.

²⁹ <<https://github.com/rogersmendonca/ETL4LOD>>. Acesso em: 8 dez. 2018.

Este agendamento pode ser feito através de um *scheduler* integrado à ferramenta ETL ou aplicações nativas do Sistema Operacional para este fim, como o Agendador de Tarefas do Windows e o *Crontab* do CentOS.

Independente do método de *scheduling* utilizado, é imprescindível o registro do que é feito a cada ciclo de execução do agendamento. Desta forma, tem-se um histórico de ações realizadas sobre um conjunto de dados, dentro de um período de tempo, sob quais condições.

4 SCHEDPINGER

Neste trabalho, foram especificados e implementados processos para extração, transformação e carga (ETL) dos dados brutos coletados do PingER, além de uma rotina de *scheduling* que permitiu automatizar e otimizar atividades que anteriormente eram realizadas de forma manual e pouco sistemática. A esta solução foi dada o nome de SchedPingER, apresentada neste capítulo.

4.1 PROCESSO ETL

O processo ETL é aplicado sobre os dados brutos do PingER, disponibilizados em arquivos TXT semi-estruturados, com seu conteúdo separado pelo caractere espaço. Para a criação desses arquivos, existem *scripts* específicos executados manualmente no SLAC, sobre os dados recebidos. Este é um processo interno, *ad hoc* e não modificado nesta proposta. Logo, quando citados os dados brutos do PingER, refere-se a estes previamente tratados e dispostos em diferentes arquivos. Na Figura 5 pode ser observada cada etapa do fluxo de tratamento dos dados e quais delas correspondem às realizadas neste trabalho.

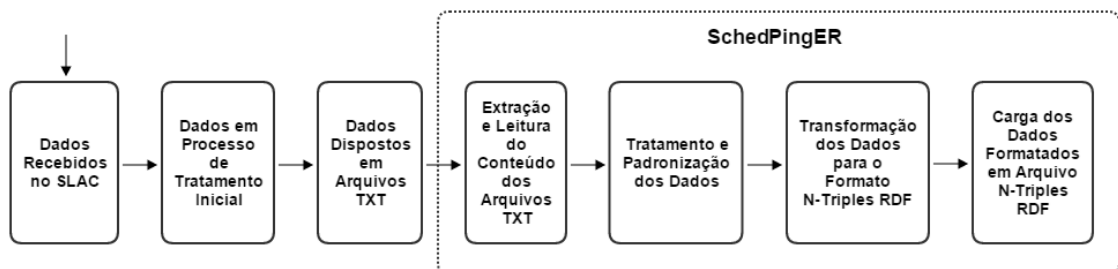


Figura 5 - Fluxo de tratamento dos dados do PingER

O conteúdo de cada um dos arquivos TXT refere-se ao valor medido por um *Ping* entre dois nós (nó fonte e nó destino) em um período de tempo. O tamanho de pacote enviado no *Ping*, em relação a qual métrica é o valor medido e a granularidade do dado são parâmetros que também descrevem uma medida PingER, como mostrado na Figura 1, na Subseção 2.1.1. Estes parâmetros, por padrão, são especificados no nome do arquivo. Assim, um arquivo nomeado “throughput-100-by-node-2015-09.txt” contém medidas na granularidade diária, para os dias de setembro de 2015, tamanho de pacote 100 *bytes*, métrica *Throughput*, para todos os pares de nós fonte e destino.

No SchedPingER, os dados brutos do PingER são extraídos dos vários arquivos TXT, tratados e transformados para o formato N-Triples RDF, segundo ontologia definida em (SOUZA, 2013). Este formato foi escolhido por facilitar o cruzamento de base de dados diferentes na *Web* de Dados, agregando valor às informações que poderão ser extraídas ao analisar tais dados.

Logo, ao final da execução da transformação dos dados brutos do PingER, cada medida é descrita por 7 triplas: uma definindo que trata-se de uma medida e outras 6 referentes a cada parâmetro que a descreve. Por exemplo, se no dia 20 de janeiro de 2016, para uma métrica qualquer “M1”, tamanho de pacote 100 *bytes*, em um *Ping* entre o nó fonte “No1” e o nó destino “No2” o valor diário medido foi “29.433”, como identificador desta medida teríamos o URI “<http://www-iepm.slac.stanford.edu/pinger/lod/resource/#No1-No2-M1-100-Time2013Jan19>” e ela seria descrita pelas triplas no modelo apresentado na Figura 6.

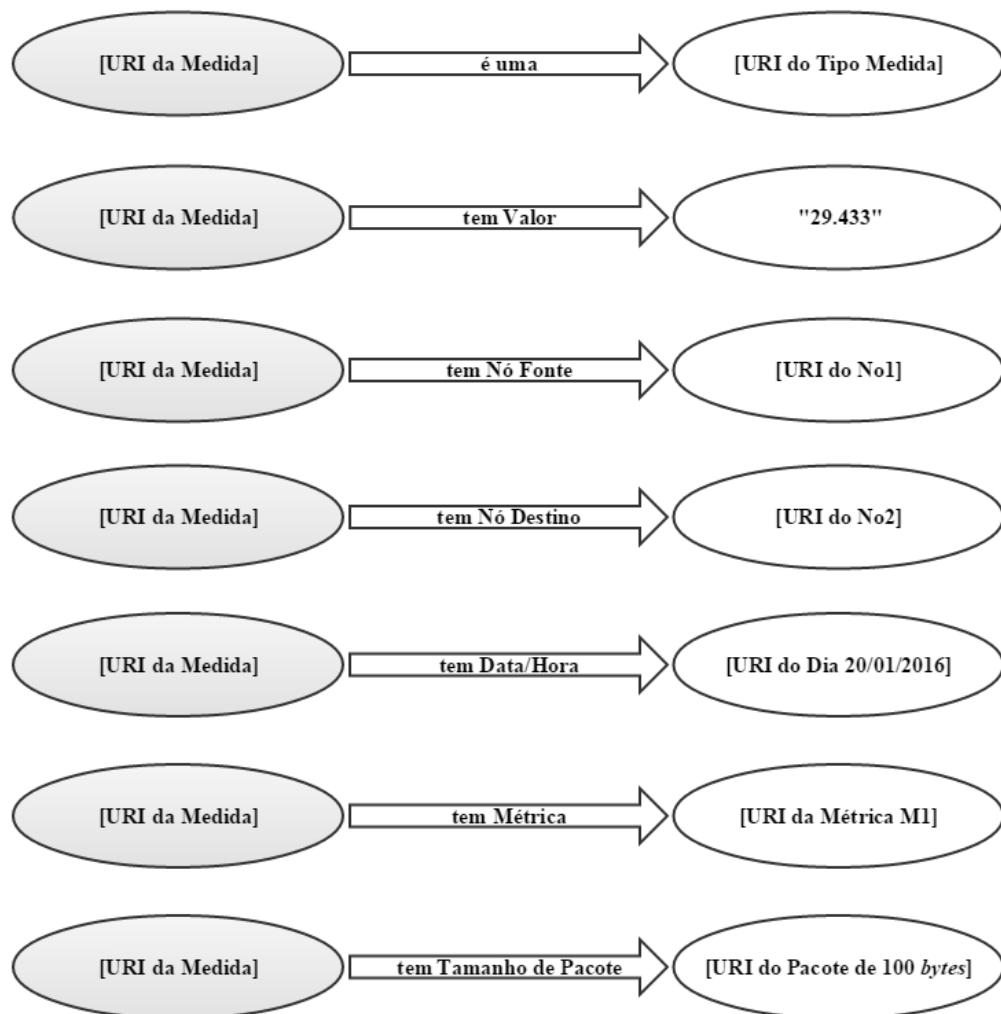


Figura 6 - Exemplo de triplas para uma medida PingER

As triplas geradas são armazenadas em um arquivo de saída com extensão N-Triples, pronto para ser carregado em um sistema de gerenciamento de dados que aceite este formato (4store, Jena, Sesame, Virtuoso, etc). No Apêndice A, um exemplo do conteúdo gerado é explicitado a fim de auxiliar a compreensão. Contudo, é importante salientar que esta transformação pode ser modificada a fim de gerar dados em diferentes formatos. Há um ponto-chave que possibilita transformar esses dados brutos para o modelo de dados relacional, por exemplo, e carregá-lo em um banco de dados SQL.

4.1.1 ETL passo-a-passo

O processo ETL dos dados do PingER é feito através de transformações implementadas na versão 4.1.0 *stable* do PDI, combinado com o *plugin* ETL4LOD, o qual apresenta *steps* que permitem manipular dados para formatá-los em LOD.

No SchedPingER temos 3 tipos de transformações: para medidas em granularidade diária, para medidas em granularidade horária e para os parâmetros que descrevem uma medida. Nos itens a seguir, são detalhadas as etapas de cada tipo de transformação.

Adicionalmente, no Apêndice B são evidenciadas as propriedades definidas em cada passo do fluxo de trabalho construído no PDI para a solução SchedPingER quando os dados de entrada estão na granularidade diária, a fim de auxiliar no entendimento e na reprodução da solução proposta. As propriedades do fluxo para dados de entrada na granularidade horária não foram explicitadas, pois são análogas às demonstradas no apêndice supracitado.

4.1.1.1 Granularidade diária

Passo 1 - *PingER Measurements TXT Input*



Figura 7 - Passo 1 da transformação para a granularidade diária

Definição das propriedades do arquivo de entrada a ser processado.

Inicialmente, a definição do caminho onde encontra-se o arquivo de entrada é realizada por passagem de parâmetro, especificado no momento da execução da transformação.

Além disto, são definidos o tipo de arquivo esperado como entrada e quais campos estão contidos nele, bem como seus respectivos tipos de dados.

Passo 2 - *Split Filenames Fields*



Figura 8 - Passo 2 da transformação para a granularidade diária

Split do nome do arquivo de entrada. Neste passo são extraídos os parâmetros para qual métrica, tamanho de pacote e data são as medidas do arquivo, através da divisão do nome do arquivo a cada caractere “-” (hífen).

Passo 3 - *Strings cut*



Figura 9 - Passo 3 da transformação para a granularidade diária

Este passo complementa o anterior, uma vez que entre o final do nome do arquivo e a extensão dele não há um caractere “-” (hífen). Nele o texto referente à extensão do arquivo é descartado, apenas o número referente à data da medida é armazenado.

Passo 4 - *Month Label Mapper*

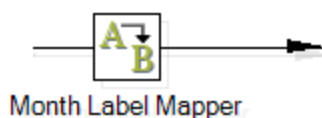


Figura 10 - Passo 4 da transformação para a granularidade diária

Correlaciona número do mês com a abreviatura de seu nome. É neste passo que é definido que o mês “01” representa janeiro, o “02” fevereiro, e assim por diante. São criados *labels*, em inglês, referentes a cada mês do ano: *Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov e Dec*.

Passo 5 - *Metric Name Mapper*

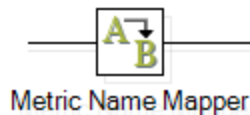


Figura 11 - Passo 5 da transformação para a granularidade diária

Ajusta o nome da métrica proveniente do arquivo de entrada, a fim de padronizar essa representação. Por exemplo, a métrica “average_rtt” passa a ser representada como “AverageRTT” e a “throughput”, como “Throughput”.

Passo 6 - *Script: Value Validation*

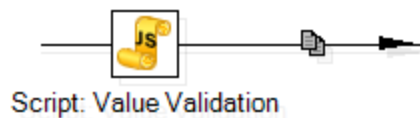


Figura 12 - Passo 6 da transformação para a granularidade diária

Script para tratar a variação da quantidade de dias em um mês. Este *script* foi implementado uma vez que meses podem ter de 28 a 31 dias. Ele confere se os valores contidos nas variáveis correspondentes às medidas dos dias 29, 30 e 31 são numéricos. Caso não sejam, o valor nulo é atribuído à variável, indicando que não há medida para o dia.

Passo 7 - *Formula: URIs*

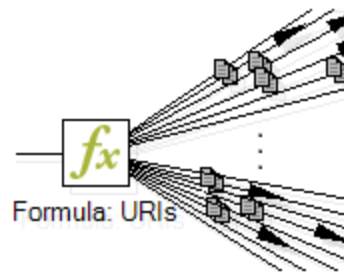


Figura 13 - Passo 7 da transformação para a granularidade diária

Descrição dos identificadores de cada recurso (URIs), segundo a ontologia definida em (SOUZA, 2013).

Passo 8 - *Filter rows*

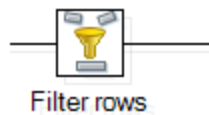


Figura 14 - Passo 8 da transformação para a granularidade diária

Confere se há medida para cada um dos 31 possíveis dias (número máximo) de um mês. Se sim, segue para o próximo passo.

Passo 9 - *Measurement Data Property Mapping / Measurement Object Property Mapping*



Figura 15 - Passo 9 da transformação para a granularidade diária

Definição de sujeito, predicado e objeto que formam uma tripla, a partir das definições de ontologia e valores obtidos nos passos anteriores.

Passo 10 - *Add constants*

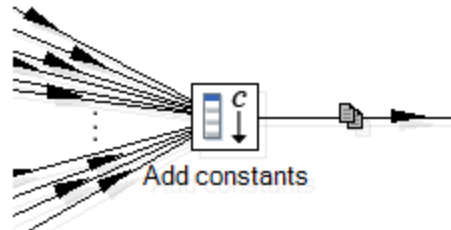


Figura 16 - Passo 10 da transformação para a granularidade diária

Neste passo são definidas constantes necessárias para a construção das triplas no formato N-Triples RDF. São elas: *scape*, representando as aspas duplas; *floatSchema*, representando o esquema definido pelo *World Wide Web Consortium (W3C)*³⁰ para indicar um valor do tipo *float* (tipo do valor da medida PingER).

Passo 11 - *Formula: DataProp / Formula: ObjectProp*

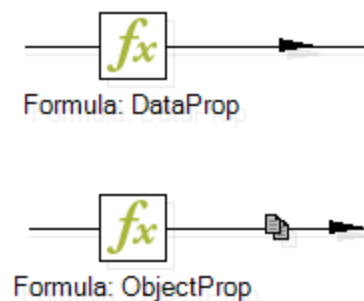


Figura 17 - Passo 11 da transformação para a granularidade diária

Construção da tripla. Neste passo os sujeitos, predicados e objetos definidos no Passo 9 são formatados no padrão N-Triples RDF.

Passo 12 - *DataProp Triples / ObjectProp Triples*

³⁰ Consórcio internacional para desenvolver padrões para a *Web*, objetivando atingir todo seu potencial. Disponível em: <<https://www.w3.org/>>. Acesso em: 8 dez. 2018.

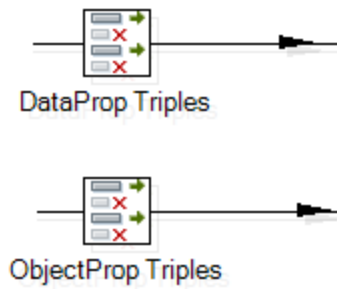


Figura 18 - Passo 12 da transformação para a granularidade diária

Seleciona a tripla construída para ser entrada no próximo passo.

Passo 13 - *Measurement Triples Output*

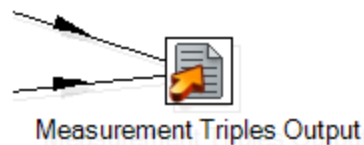


Figura 19 - Passo 13 da transformação para a granularidade diária

Gera arquivo de saída, em formato N-Triples RDF, com todas as triplas geradas no processo.

4.1.1.2 Granularidade horária

A transformação para medidas em granularidade horária segue o mesmo raciocínio que a transformação para medidas em granularidade diária, com algumas adaptações:

- A granularidade em vez de ser definida por ano e mês é definida por ano, mês e dia.
- Não é mais necessário o Passo 6, do *script* de validação, uma vez que todos os dias têm 24 horas de duração.
- O Passo 8 confere se há medida para cada uma das 24 horas de um dia.

4.1.1.3 Parâmetros descritivos de medida

Além das transformações das medidas PingER, secundariamente, temos uma transformação geradora de triplas que definem os parâmetros e as informações que auxiliam na descrição de uma medida. Nela, são produzidas triplas que especificam atributos dos nós considerados no projeto; dos continentes, países, estados e cidades em que localizam-se estes nós; das datas e horários dos *Pings*; dos tamanhos de pacote transmitidos; das métricas de rede e das unidades de medida utilizadas.

Essas triplas também se apresentam no formato N-Triples RDF, segundo ontologia definida em (SOUZA, 2013). Esta baseia-se na ontologia MOMENT³¹ e utiliza ontologias *GeoNames*³² (para descrição de localidades) e *Time*³³ (para descrição de tempo), do W3C.

Exceto pelo fluxo ETL correspondente às triplas de descrição temporal, todos os outros recebem como entrada arquivos no formato CSV. Estes diversos fluxos contidos na transformação serão explicitados a seguir.

Fluxo de descrição dos nós

Como mostrado na Figura 20, o arquivo de saída correspondente às informações sobre os nós considerados no PingER resultam do processamento de um arquivo de entrada chamado *Node Details*. Nele estão contidos os seguintes dados sobre os nós fonte e destino: ID; latitude e longitude (coordenadas geográficas); IDs e nomes do país, do estado e da cidade onde estão localizados; endereço IP (*Internet Protocol*); nome e *nickname* do nó; nome da localidade (por exemplo, a universidade em que o nó está); descrição da localização e outras propriedades de controle interno.

Os dados supracitados são obtidos a partir do processamento do conteúdo disponibilizado publicamente pela equipe do PingER através de uma página *Web*³⁴, onde as informações estão dispostas de forma estruturada, permitindo a leitura e o tratamento prévio em que os dados são convertidos para o formato CSV.

³¹ <http://www.salzburgresearch.at/en/projekt/moment_en/>. Acesso em: 8 dez. 2018.

³² <<http://www.geonames.org/ontology/documentation.html>>. Acesso em: 8 dez. 2018.

³³ <<https://www.w3.org/TR/owl-time/>>. Acesso em: 8 dez. 2018.

³⁴ <<http://www-iep.mslac.stanford.edu/pinger/pingerworld/nodes.cf>>. Acesso em: 8 dez. 2018.

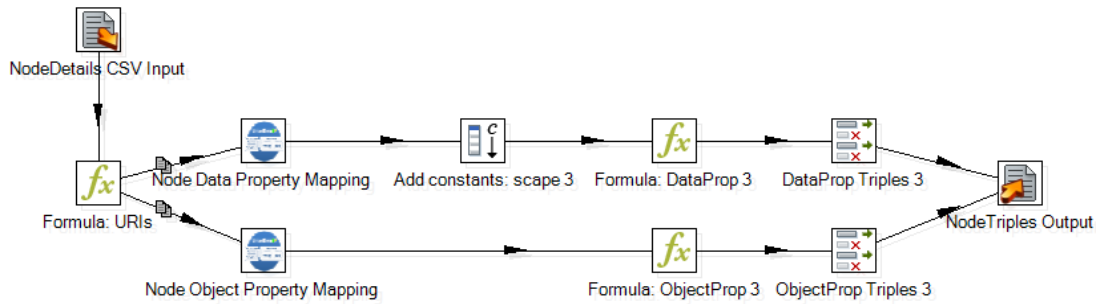


Figura 20 - Representação do fluxo ETL de descrição dos nós

No início do fluxo de descrição dos nós, cada um dos campos do arquivo CSV é lido. Após, são gerados os padrões de representação dos dados. Formula-se o URI do nó, a fim de identificá-lo; e URIs do seu país, estado e cidade, utilizando os campos contendo seus IDs. Desta forma, é possível um nível de detalhamento da localidade, a partir do cruzamento de triplas geradas em outros fluxos da mesma transformação.

O sujeito, o predicado e o objeto que formam uma tripla são definidos, considerando todos os campos do arquivo de entrada. Há um passo de construção dessas triplas, totalizando 19 para descrever cada nó. Por fim, estas são carregadas em um arquivo de saída em formato N-Triples RDF.

Fluxo de descrição dos continentes

Como mostrado na Figura 21, os dados dos continentes são retirados de um arquivo de entrada em formato CSV, elaborado manualmente, contendo o ID e o nome de cada um dos sete continentes; seus *links* no *GeoNames* e na *DBpedia*; seu código/sigla composto por duas letras.

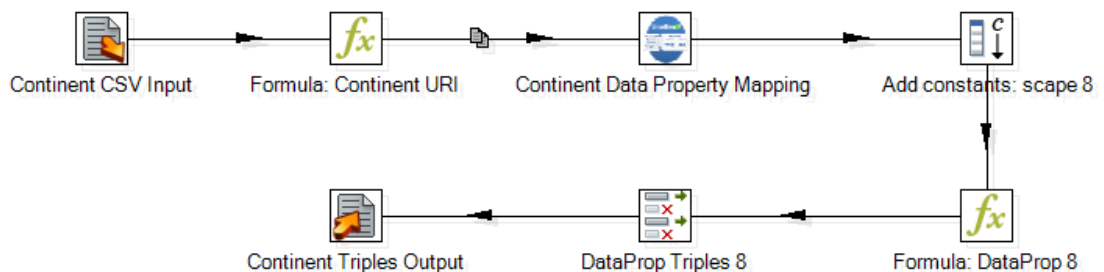


Figura 21 - Representação do fluxo ETL de descrição dos continentes

Nesse fluxo, é necessária a definição de apenas um URI: o do continente. A definição de sujeito, predicado e objeto que compõem uma tripla é realizada, gerando 7 triplas para cada continente. Estas são carregadas em um arquivo de saída em formato N-Triples RDF.

Fluxo de descrição dos países

Como mostrado na Figura 22, as triplas correspondentes aos países são geradas a partir dos dados contidos em um arquivo de entrada em formato CSV, confeccionado pela aplicação *getCountries()* disponível no *GitHub*³⁵, em que determinados dados sobre todos os países do globo são coletados do *GeoNames* através de uma interface amigável. Nesta aplicação são obtidos o ID do país no *GeoNames*; seu nome; seu código/sigla composto por duas letras; sua unidade monetária; sua população; sua capital; o nome e o código/sigla do continente do qual faz parte; sua área em quilômetros quadrados; seus idiomas oficiais.

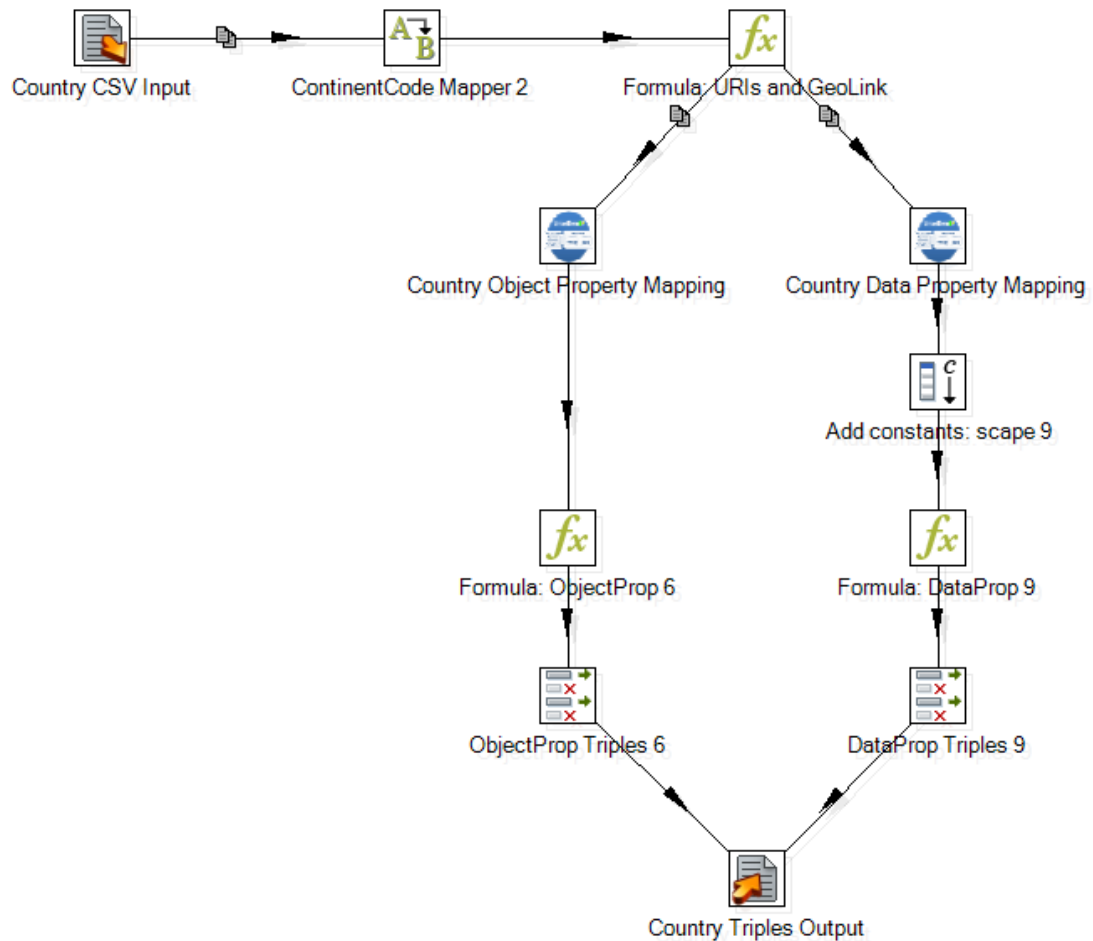


Figura 22 - Representação do fluxo ETL de descrição dos países

O fluxo de transformação inicia-se com a leitura de todo conteúdo do arquivo de entrada. A seguir, é feito um mapeamento para saber, a partir da sigla do continente, qual seu

³⁵ <<http://peric.github.io/GetCountries/>>. Acesso em: 8 dez. 2018.

ID no *GeoNames*. Assim, é possível o cruzamento de dados com as triplas de descrição dos continentes, agregando informações.

Uma vez conhecido o ID do continente, há o passo de definição do URI do país e de seu continente, bem como a do endereço da página do país no *GeoNames*. Após, também define-se sujeito, predicado e objeto formadores de uma tripla. Cada país é descrito por 12 triplas. Estas são geradas pelo fluxo e carregadas em um arquivo de saída em formato N-Triples RDF.

Fluxo de descrição dos estados e cidades

Como mostrado na Figura 23, os dados referentes aos estados e às cidades – nos quais nós utilizados no PingER localizam-se – são obtidos através do processamento de um arquivo de entrada no formato CSV, elaborado nos trabalhos associados ao grupo de pesquisa MultiLod. Nele são dispostos latitude e longitude das cidades (coordenadas geográficas); ID do país, do estado e da cidade no *GeoNames*; nome do país, do estado e da cidade; sigla do continente em que a cidade está inserida; *links* da cidade no *GeoNames*, na *Wikipedia* e na *DBpedia*; entre outros.

Primeiro, cada um dos campos do arquivo de entrada é lido. Após, ocorre o mapeamento do nome e do ID no *GeoNames* do continente a partir de sua sigla. São definidos os URIs correspondentes ao continente, ao país, ao estado e à cidade em que nós do PingER são encontrados.

Este fluxo de transformação subdivide-se em dois, uma vez que ele gera duas saídas diferentes: uma para os estados, outra para as cidades consideradas. Assim, nos passos seguintes, são definidos o sujeito, o predicado e o objeto formadores das triplas, de forma ramificada. Estas são construídas e carregadas em seus respectivos arquivos de saída em formato N-Triples RDF.

Para descrever um estado são utilizadas 9 triplas, enquanto que, para uma cidade, são utilizadas 22 triplas.

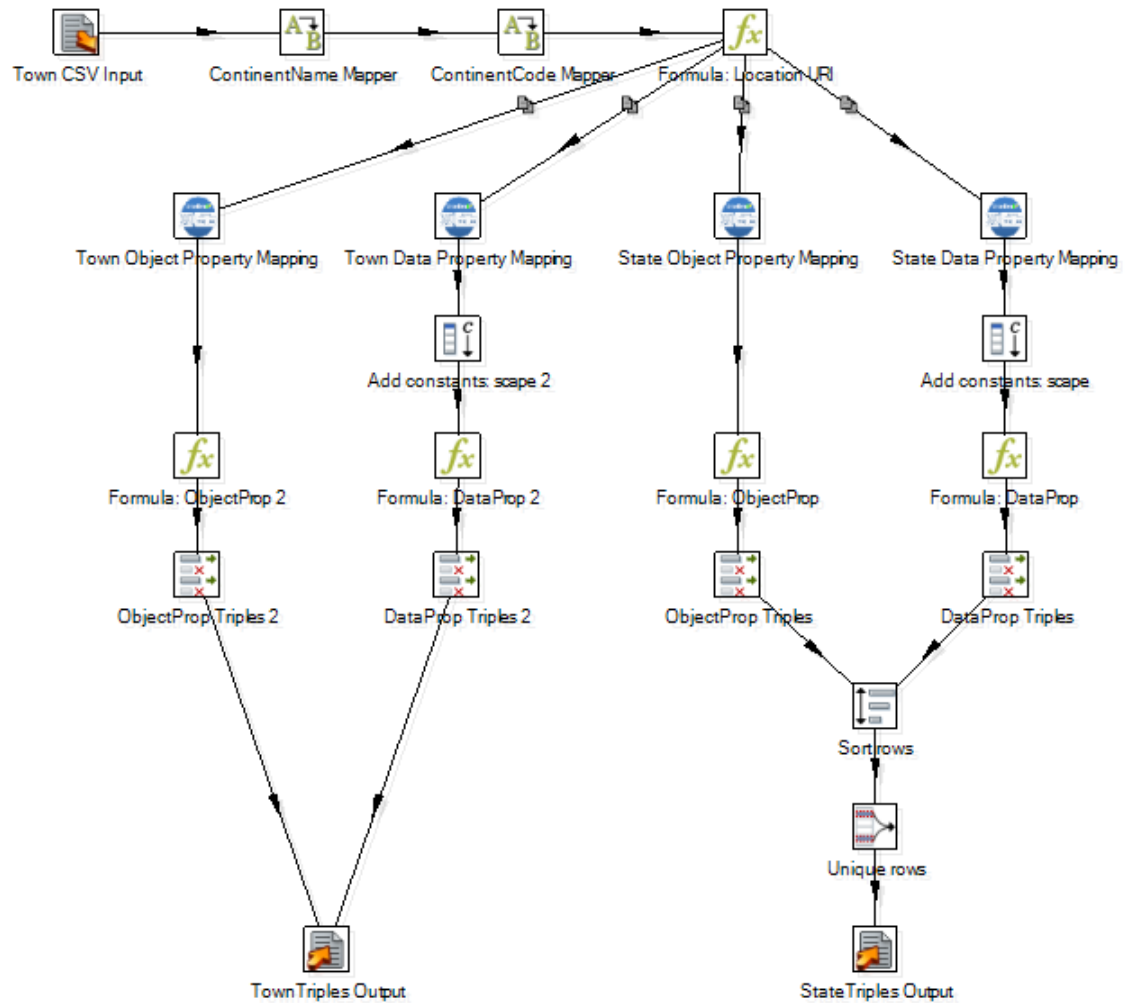


Figura 23 - Representação do fluxo ETL de descrição dos estados e cidades

Fluxo de descrição das métricas

Como mostrado na Figura 24, o fluxo de descrição das métricas utilizadas pelo PingER recebe como entrada um arquivo elaborado manualmente, em formato CSV, contendo ID, nome e unidade de medida padrão da métrica.

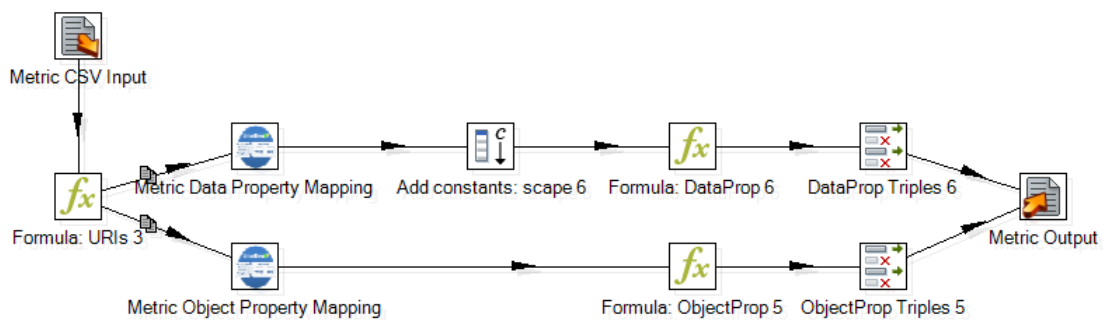


Figura 24 - Representação do fluxo ETL de descrição das métricas

Primeiro, o arquivo de entrada é lido. São definidos os URIs da métrica e da sua unidade de medida padrão. O sujeito, o predicado e o objeto que formam uma tripla são indicados no passo seguinte. Para cada métrica, há 3 triplas descrevendo-a. Estas são carregadas em um arquivo de saída em formato N-Triples RDF.

Fluxo de descrição dos tamanhos de pacote

Como mostrado na Figura 25, os dados referentes aos tamanhos de pacote utilizados no PingER são apresentados em um arquivo de entrada em formato CSV, elaborado de forma manual. Ele contém o valor e a unidade de medida correspondentes.

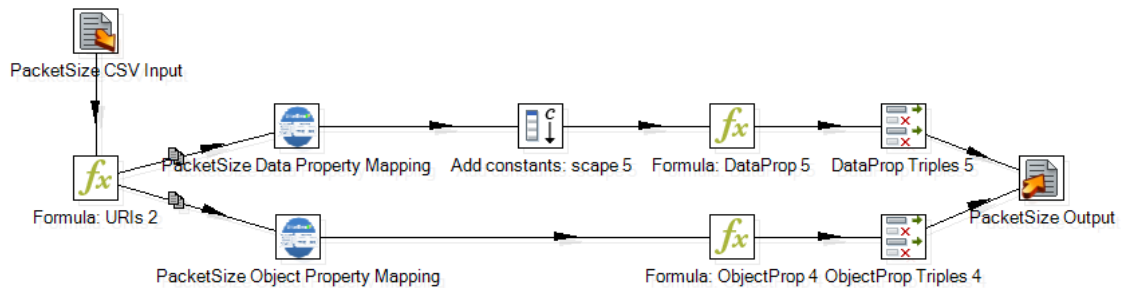


Figura 25 - Representação do fluxo ETL de descrição dos tamanhos de pacote

O fluxo de transformação inicia-se com a leitura do arquivo de entrada. Após, são definidos os URIs do valor numérico que representa o tamanho de pacote e de sua unidade de medida padrão. O sujeito, o predicado e o objeto que formam uma tripla são indicados no passo seguinte. São geradas apenas 6 triplas neste fluxo. Estas são carregadas em um arquivo de saída em formato N-Triples RDF.

Fluxo de descrição das unidades de medida

Como mostrado na Figura 26, há um arquivo de entrada em formato CSV, contendo o nome e o símbolo das 5 unidades de medida utilizadas no PingER. Este arquivo foi elaborado manualmente, a partir das unidades de medida descritas na definição de cada métrica considerada no projeto PingER.

Os dados são lidos do arquivo de entrada, o URI da unidade é definido e as 2 triplas necessárias para descrever cada unidade de medida são construídas, através da definição de sujeito, predicado e objeto. Estas são carregadas em um arquivo de saída em formato N-Triples RDF.

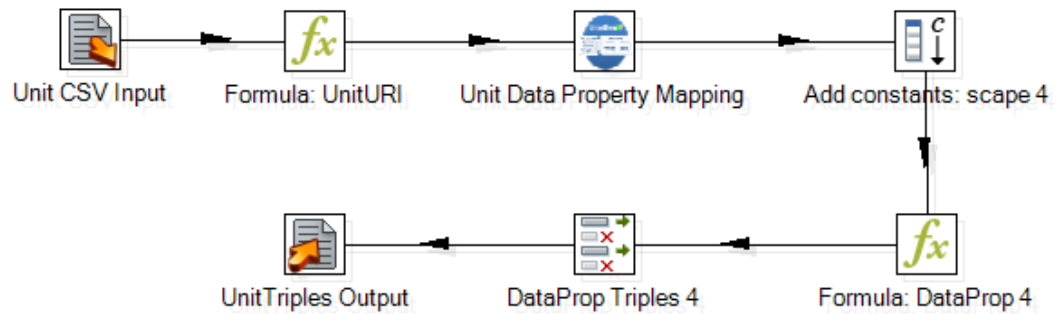


Figura 26 - Representação do fluxo ETL de descrição das unidades de medida

Fluxo de descrição temporal

Como mostrado na Figura 27, o fluxo gerador de triplas correspondentes à descrição temporal é o único que não recebe um arquivo de entrada para leitura. A transformação inicia-se com a definição das datas inicial e final a serem consideradas.

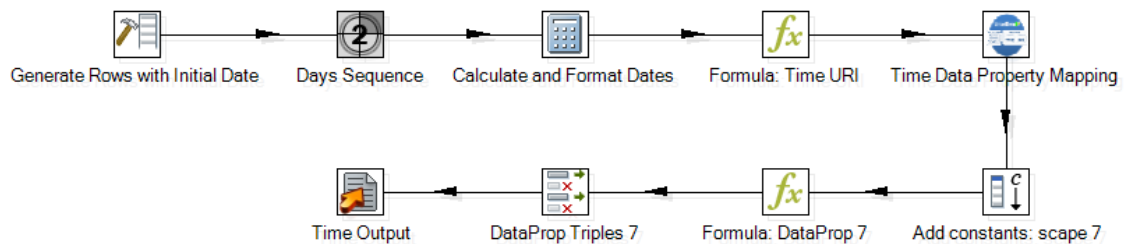


Figura 27 - Representação do fluxo ETL de descrição temporal

A definição da data inicial ocorre explicitamente, informando um valor no formato escolhido. Enquanto que a data final é sabida através da definição de um limite para a quantidade de datas a serem geradas a partir da data inicial. Portanto, se quisermos gerar triplas para todos os dias de um ano não bissexto, por exemplo, deve-se definir como data inicial o primeiro dia do ano e, como limite, o valor 365 representando cada dia transcorrido.

A partir desta definição, são realizados os cálculos em que serão conhecidas as propriedades para o período considerado. São definidos dois URIs: um indicando granularidade, outro explicitando o momento temporal de acordo com esta granularidade. Assim, para a granularidade diária, por exemplo, será definido um URI informando a granularidade diária e outro representando um dia específico.

Após, o sujeito, o predicado e o objeto que formam uma tripla são indicados. São geradas até 9 triplas para cada momento em que pode existir uma medida PingER. Estas são carregadas em um arquivo de saída em formato N-Triples RDF.

4.2 SCHEDULING

O PDI apresenta perspectiva integrada de *scheduling* somente na edição *Enterprise*, versão sob licença comercial. Por isso, o agendamento da execução dos fluxos ETL foram feitos através de serviços oferecidos pelos Sistemas Operacionais Windows e CentOS.

Assim, foram implementados arquivos executáveis nos quais são descritos os comandos necessários para a execução de cada transformação no PDI. Nas Figuras 28 e 29, são exibidos os conteúdos dos arquivos de *scheduling* para as transformações de granularidade diária. Os dois têm a mesma função, diferenciando-se apenas pelos comandos característicos de cada Sistema Operacional.

```

1 set hr=%time:~0,2%
2 set hr=%hr: =0%
3 set mainPath=E:\CRBD\PingERTriplification
4 set inputPath=%mainPath%\Input\Fact\Daily
5 set transformationPath=%mainPath%\PingERFact_Daily.ktr
6 cd /d %inputPath%
7 for /r %i in (*.txt) do %mainPath%\data-integration4.1\Pan.bat /file:%transformationPath%
  -param:PingERFile=%i" /level:Basic >> "%mainPath%\logs\PingERFactDailyTransf_
  %date:~6,4%%date:~3,2%%date:~0,2%%hr%%time:~3,2%%time:~6,2%.log"
8 exit

```

Figura 28 - Conteúdo do arquivo executável para scheduling no Windows

```

1 #!/bin/bash
2 mainPath="/home/geral/PingERTriplification"
3 inputPath="$mainPath/Input/Daily"
4 transformationPath="$mainPath/PingERFact_Daily.ktr"
5 cd $inputPath
6 for f in *.txt ;
7 do
8     $mainPath/data-integration4.1/pan.sh -file=$transformationPath -param:PingERFile=
9     $inputPath/$f -level=Basic >> $mainPath/logs/PingERFactDailyTransf_$(date +%Y%m%d%H%M%S).log
10 done
11 exit

```

Figura 29 - Conteúdo do arquivo executável para scheduling no CentOS

Em cada executável, temos a indicação de qual diretório contém os arquivos TXT de entrada (linha 4 da Figura 28 e linha 3 da Figura 29) e o caminho da transformação a ser executada através do componente *Pan* do PDI (linha 5 da Figura 28 e linha 4 da Figura 29).

Os arquivos de *log* são gerados de forma padronizada, para manter um registro de todos os passos que ocorrem a cada vez que uma transformação é executada. Deste modo, tem-se o controle de quais arquivos TXT de entrada foram considerados, o que ocorreu em cada etapa da transformação, quantas triplas foram geradas ao final e se houve algum erro durante a execução. Portanto, toda vez que uma transformação é executada, um arquivo de *log* é gerado com o seguinte padrão de nome: tipo da transformação, seguido de um código único composto pelo ano, mês, dia, hora, minuto e segundo que ela foi realizada.

A geração deste código identificador é feita a partir da data e da hora do sistema, portanto seguem a formatação local. Por esta razão, no arquivo executável do Windows, um tratamento é feito para o valor referente ao campo hora. Quando a hora do sistema é de 0 a 9h, o valor retornado é um caractere espaço mais o número. Então, uma adaptação é feita para manter o padrão de dois dígitos (linhas 1 e 2 da Figura 28).

Caso o formato da data do sistema seja modificado, ou seja, a ordem de apresentação de dia, mês e ano seja alterada, também serão necessárias adaptações no conteúdo do executável, uma vez que, no Windows, cada um desses campos são posições dentro de uma cadeia de caracteres correspondente à data completa.

Para o sistema CentOS, não são necessárias adaptações específicas para formatos de data e hora, dado que cada valor que compõe data e hora completas do sistema é representado por uma variável distinta e o campo hora apresenta sempre dois dígitos.

Uma vez preparados os executáveis, é possível agendá-los para executar em períodos pré-determinados, sem que haja a necessidade de alguém iniciando e supervisionando cada transformação. Assim, o tratamento dos dados do PingER é feito de forma sistematizada e controlada, com uma frequência definida.

No Windows, o *scheduling* pode ser feito tanto por linha de comando, através do comando *at* e seus parâmetros, como através do Agendador de Tarefas – interface gráfica de usuário disponível a partir da versão Windows 2000, acessível pelo Painel de Controle. No SchedPingER, utilizamos esta segunda opção. Basta abrir a interface do agendador e selecionar a ação de criar tarefa básica. Uma janela é aberta e 4 etapas seguem:

- Nomear e descrever a tarefa a ser agendada.
- Definir a frequência em que a tarefa será realizada. Nesta etapa define-se o início e o critério de execução, podendo ser por periodicidade ou a partir do registro da ocorrência de um evento específico.
- Definir a ação a ser realizada. Neste caso, definir que a ação é iniciar um programa e indicar o caminho para o arquivo executável previamente implementado.
- Conferir o resumo da tarefa contendo as escolhas feitas nas etapas anteriores e concluir o agendamento.

No CentOS, o *scheduling* é feito por linha de comando, através do comando *crontab*. Este, seguido do parâmetro *-e*, permite editar (ou criar, caso não exista) o arquivo no qual são

especificadas as ações a serem realizadas pelo *Cron*³⁶ do sistema, em um período e frequência programados.

Neste arquivo, o agendamento de uma ação corresponde a uma linha composta por 6 campos, na respectiva ordem: minuto, hora, dia do mês, mês, dia da semana e ação a ser executada. Cada um deles podem receber os seguintes valores:

- Minuto: números de 0 a 59.
- Hora: números de 0 a 23.
- Dia do mês: números de 1 a 31.
- Mês: números de 1 a 12, correspondendo aos meses de janeiro a dezembro.
- Dia da semana: números de 0 a 6, correspondendo aos dias de domingo a sábado.
- Ação a ser executada: um comando válido do Sistema Operacional.

Nos campos numéricos é possível especificar um ou mais valores e intervalos. Para informar mais de um valor para o mesmo campo, basta que estes sejam separados por vírgula (.). Para intervalos, basta separar os números de início e fim com um hífen (-). Para considerar quaisquer valores possíveis, basta atribuir asterisco (*) ao campo.

No SchedPingER, a transformação dos dados na granularidade horária pode ter sua execução agendada para todos os dias, às 2h, por exemplo. Assim, uma vez que o histórico de medidas PingER seja processado, sempre serão tratados os dados horários obtidos no dia anterior, a fim de manter a base constantemente atualizada. No *crontab*, esta frequência de execução é definida pela linha “0 2 * * * [caminho para o arquivo executável de *scheduling* previamente criado]”.

³⁶ O *Cron* é um serviço do sistema operacional CentOS, encarregado de verificar se há alguma ação a ser realizada e executá-la conforme programada.

5 EXPERIMENTAÇÃO

Neste capítulo, analisamos a solução SchedPingER a partir do estudo sobre a transformação de amostras dos dados, abrangendo as granularidades horária e diária do projeto PingER.

5.1 OBJETIVO

Conforme abordado anteriormente, a proposta SchedPingER apresentada neste trabalho visa sistematizar uma rotina de extração, transformação e carga dos dados de qualidade de rede ao redor do mundo, obtidos através do projeto PingER desde 1998 até os dias atuais.

Portanto, foram definidos conjuntos de arquivos PingER como insumo para testes de desempenho dos fluxos de transformação construídos na versão 4.1.0 *stable* do PDI, de forma a verificar a viabilidade da aplicação para o processamento do histórico de dados do projeto e de novos valores medidos, conforme estes forem gerados.

5.2 CONFIGURAÇÃO DA EXPERIMENTAÇÃO

Uma vez que os dados do PingER são apresentados em diversas granularidades, restringimos nossa solução e experimentação para os cenários em que os dados são horários ou diários.

Conforme a configuração dos arquivos disponibilizados via FTP público, os dados horários são agrupados diariamente, em arquivo contendo medidas das 24 horas de determinado dia; enquanto os diários são agrupados mensalmente, em arquivo contendo medidas para cada um dos dias de determinado mês (variando entre 28 e 31 dias).

Portanto, para os testes iniciais, partimos do menor volume possível de dados, sendo um dia para análise da granularidade horária, e um mês para a granularidade diária. Cabe ressaltar que para os dados horários são consideradas todas as 16 métricas do projeto, enquanto que para os diários são consideradas apenas 5 métricas devido à restrição de disponibilidade dos arquivos.

Após, o insumo para cada uma das granularidades cresce gradativamente, representando uma periodicidade maior na entrada de dados a fim de avaliar o desempenho e

a viabilidade de tratamento do histórico de medidas PingER. Assim, foram estabelecidos os seguintes períodos de dados para experimentação:

- Granularidade horária:
 - 1 dia (16 arquivos).
 - 1 semana (112 arquivos).
 - 1 mês (480 arquivos).
- Granularidade diária:
 - 1 mês (5 arquivos).
 - 4 meses (20 arquivos).
 - 1 ano (60 arquivos).
 - 41 meses (196 arquivos).

Todos os testes foram realizados em um computador com sistema operacional *Microsoft Windows 7 Ultimate* de 64 bits; processador *Intel Core i3-2330M*, com frequência 2,20 gigahertz, 2 núcleos de processamento; e memória RAM de 4 gigabytes. Alguns dos testes listados também foram realizados em um computador com sistema operacional *CentOS Release 6.5* de 64 bits; processador *Intel Core i7-4770*, com frequência 3,40 gigahertz, 8 núcleos de processamento; e memória RAM de 8 gigabytes. No entanto, não foram observadas diferenças significativas em relação ao tempo de processamento dos dados, de forma que os resultados apresentados referem-se aos realizados no ambiente Windows. Os fluxos de transformação de dados foram executados através do componente *Pan* da versão 4.1.0 *stable* do PDI, acrescido do conjunto de *plugins* ETL4LOD.

5.3 RESULTADOS

5.3.1 Granularidade horária

Para a granularidade horária, o menor conjunto de testes refere-se a 16 arquivos, representando cada métrica considerada no projeto PingER para um dia do ano. O dia 1 de julho de 2016 foi escolhido arbitrariamente, em que as medições PingER somam 1.782.250 valores, processados em 9min22s, resultando em 12.475.750 triplas.

Após, verificamos o desempenho da solução ao acumular arquivos de determinados períodos para execução do fluxo ETL. Portanto, o segundo teste foi realizado a partir de uma

entrada contendo 112 arquivos, os quais representam cada métrica vista no projeto PingER, para o período de uma semana, entre os dias 1 e 7 de maio de 2016, totalizando 5.549.041 medidas PingER, processadas em 55min25s, resultando em 38.843.287 triplas.

Por último, como insumo para a solução, fornecemos 480 arquivos que representam as medidas para cada métrica PingER de um mês do projeto – neste caso, os trinta dias entre o dia 1 e 30 de abril de 2016 – totalizando 54.452.558 de valores na entrada, processados em 4h34min10s, resultando na geração de 381.167.906 triplas.

Todos os resultados citados acima, encontram-se resumidos na Tabela 1.

Tabela 1 - Resultado da experimentação com dados de granularidade horária

DADOS GERAIS	1 DIA (01/07/2016)	1 SEMANA (01-07/05/2016)	1 MÊS (04/2016)
Quantidade de Arquivos de Entrada	16	112	480
Quantidade de Medidas de Entrada	1.782.250	5.549.041	54.452.558
Quantidade de Tripas Geradas	12.475.750	38.843.287	381.167.906
Tamanho do Arquivo de Saída (em <i>gigabytes</i>)	3,2	10,1	98,8
Tempo de Processamento (hh:mm:ss)	00:09:22	00:55:25	04:34:10

5.3.2 Granularidade diária

Para a granularidade diária, encontramos uma limitação de disponibilidade de arquivos PingER para apenas 5 das 16 métricas consideradas no projeto. Portanto, todos os testes realizados consideram apenas as seguintes métricas: *Mean Opinion Score* (MOS), *Maximum Round Trip Time* (*Maximum* RTT), *Packet Loss*, *Unreachability* e *Ping Unpredictability*.

Como menor conjunto de teste, especificamos o correspondente a um mês de medições PingER – em março de 2015 – totalizando 5 arquivos e 1.170.788 medidas, que foram processadas em 1min54s, resultando na geração de 8.195.516 triplas.

Após, verificamos o desempenho do SchedPingER ao processar 20 arquivos na granularidade diária, que referem-se a 4 meses de medições – entre setembro e dezembro de 2015 – representando 5.104.346, processadas em 8min11s, resultando em 35.730.422 triplas.

A partir de 60 arquivos de entrada, correspondentes a um ano de medidas PingER durante todo o ano de 2013, totalizando 12.757.647 de valores, um terceiro teste foi realizado. Neste, todo o fluxo ETL durou 51min02s de processamento, somando 89.303.529 triplas ao resultado.

Por último, consideramos as medidas PingER de 41 meses – entre dezembro de 2011 e abril de 2015 – distribuídas em 196 arquivos, totalizando 43.420.095 valores na entrada, processados em 1h11min58s, resultando em 303.940.665 triplas.

Todos os resultados citados acima, encontram-se resumidos na Tabela 2.

Tabela 2 - Resultado da experimentação com dados de granularidade diária

DADOS GERAIS	1 MÊS (03/2015)	4 MESES (09-12/2015)	1 ANO (2013)	41 MESES (12/2011-04/2015)
Quantidade de Arquivos de Entrada	5	20	60	196
Quantidade de Medidas de Entrada	1.170.788	5.104.346	12.757.647	43.420.095
Quantidade de Tripas Geradas	8.195.516	35.730.422	89.303.529	303.940.665
Tamanho do Arquivo de Saída (em <i>gigabytes</i>)	2,1	9,1	22,7	77,3
Tempo de Processamento (hh:mm:ss)	00:01:54	00:08:11	00:51:02	01:11:58

5.4 ANÁLISE DOS RESULTADOS

A partir dos resultados sintetizados nas Tabelas 1 e 2, conclui-se que um arquivo PingER com medidas na granularidade horária, para uma métrica específica, contendo, em média, 101.618 valores obtidos através dos *Pings*, leva em torno de 33,5 segundos para percorrer o fluxo ETL sugerido no SchedPingER. Analogamente, para a granularidade diária, a média de valores registrados aumenta para 222.252, mas o tempo de processamento reduz para 28,4 segundos.

Este comportamento ocorre pois, além da duração do processo de transformação dos dados originais para LOD, há uma fração de segundos a ser considerada entre as execuções para diferentes arquivos de entrada, uma vez que o componente *Pan* do PDI é executado uma vez para cada um deles. Assim, o tempo total tende a ser maior conforme o aumento da quantidade de arquivos. Logo, para a granularidade horária, em que temos um arquivo para cada dia do período selecionado para teste, a média de duração é maior embora a quantidade de valores medidos não o seja.

Para a granularidade diária, temos um arquivo para cada métrica PingER por mês. Portanto, ao final de 2018, espera-se que a quantidade total seja de 4.032 arquivos, tornando a transformação de todo o histórico desses dados viável, pois levaria um pouco mais de um dia. No entanto, esta aplicação torna-se extremamente custosa para o histórico completo de medidas horárias, uma vez que espera-se 122.720 arquivos com esta caracterização. Caso todo o histórico de dados horários fosse processado ao final de 2018, seriam necessários aproximadamente 48 dias até o fim do processamento.

Logo, conclui-se que a hipótese de implantação de uma rotina de transformação diária ou mensal, conforme os dados obtidos através dos *Pings* são consolidados, é viável desde que observados os seguintes pontos: para a hipótese de transformação de todo o histórico de dados do projeto, há viabilidade quando consideramos a granularidade diária do projeto. Quando considerados os dados horários, o ideal é que seja limitado um período inicial de amostra, para que a partir dele, todos os dados posteriores sejam gradativamente transformados.

6 CONCLUSÃO

Neste trabalho, a partir de conceitos de Dados Abertos Conectados (LOD – *Linked Open Data*), Processo de Triplificação de Dados, Extração, Transformação e Carga de dados (ETL – *Extract, Transform and Load*) e *Scheduling*, foi desenvolvida uma solução para tratamento dos dados de medidas de qualidade de rede, coletados em escala mundial através do projeto PingER, iniciado em 1998, no *Stanford Linear Accelerator Center* (SLAC), laboratório operado pela Universidade de Stanford.

A necessidade da implementação desta solução se deu uma vez que, embora haja estudos sobre os dados coletados no decorrer dos 20 anos do projeto supracitado, ainda não existia uma solução que padronizasse e facilitasse o tratamento e a carga dos valores de medidas obtidos. Estes, encontram-se disponibilizados via FTP público, em diversos arquivos com extensão TXT, o que dificulta sua manipulação e a extração de informações relevantes a partir da análise e do relacionamento dos dados com outras bases.

Portanto, sugerimos a utilização do fluxo de trabalho elaborado na ferramenta ETL *Pentaho Data Integration* (PDI – também conhecida como *Kettle*), versão 4.1.0 *stable*, acrescida do conjunto de *plugins* ETL4LOD, desenvolvido por pesquisadores do GRECO/UFRJ, em que cada arquivo gerado no projeto PingER é fornecido como entrada para uma rotina automática de extração e de transformação dos dados em triplas de formato N-Triples RDF, prontas para serem carregadas em banco de dados de preferência. Uma vez definida, esta carga em banco pode ser incluída como um passo do próprio fluxo de trabalho do PDI.

O formato em triplas RDF foi escolhido devido a possibilidade de melhor integração entre os dados do PingER e os oriundos de fontes diversas, através de uma correlação explícita utilizando o conceito de LOD, a fim de identificar eventos, situações ou locais críticos que precisam de um bom desempenho de rede e de acesso à informação.

Para verificação da aplicabilidade deste trabalho, denominado SchedPingER, foram realizados experimentos com dados PingER consolidados tanto diariamente, quanto mensalmente. Isto é, dados nas granularidades horária e diária, respectivamente. Assim, fornecemos os arquivos de entrada considerando diferentes períodos de tempo, como: medidas de um dia, de uma semana, de um mês, de um ano, etc.

No experimento de granularidade horária, observamos que o tratamento é viável para um histórico limitado de dados e para os novos conforme eles surgem. Enquanto que, no experimento de granularidade diária, esta viabilidade se dá para todo o histórico de dados e

para os futuros. Este comportamento ocorre pois a quantidade de arquivos que são fornecidos como entrada onera o tempo de processamento do fluxo. Como, para a granularidade horária, há um arquivo para cada dia do ano, este volume de entrada torna-se maior com o passar do tempo. Para a granularidade diária, os dados são apresentados em arquivos mensais consolidados, fazendo com que o número total de arquivos tenda a ser menor.

Esta dificuldade sobre o tempo de processamento ocorre pois o *Pan*, componente do PDI que executa as transformações através de linhas de comando, precisa ser iniciado uma vez para cada arquivo de entrada, através de uma rotina automática de *scheduling*. Isto poderia ser mitigado ao utilizar versões mais recentes da ferramenta, pois a partir da versão 5.1.0 foram implementados novos *steps*, como o *Transformation Executor*, o qual permite criar dentro do próprio fluxo de trabalho um passo determinando a execução de uma transformação específica do PDI uma vez para cada linha de entrada, que, neste caso, seria o caminho para cada um dos arquivos PingER.

Contudo, no SchedPingER, ficamos limitados à versão 4.1.0 *stable* do PDI devido à utilização do ETL4LOD, visando à geração de dados em formato de triplas, uma vez que a atualização deste conjunto de *plugins* demandaria um novo trabalho para entendimento quanto ao desenvolvimento de soluções integradas ao PDI e a sua realização, de fato. Este ponto vem sendo abordado e implementado atualmente no Trabalho de Conclusão de Curso (TCC) do aluno João Curcio, concluinte do Bacharelado em Ciência da Computação pela UFRJ, com o objetivo de atualizar e documentar as funcionalidades existentes no ETL4LOD, bem como adicionar novas, como as relacionadas ao gerenciamento e busca de ontologias e vocabulários a serem utilizados em um processo de triplificação.

Por fim, como ganho do presente trabalho, além da possibilidade de transformar as medidas PingER para N-Triples RDF de forma sistematizada a fim de facilitar sua manipulação e manutenção, a partir do fluxo criado é possível adaptá-lo para que sejam gerados dados de saída em outros formatos. Pois, a solução foi construída com um ponto-chave em que o fluxo no PDI pode ser modificado para o formato desejado de um modo mais intuitivo do que por linhas de código. Logo, funcionalidades contidas em outras versões da ferramenta, como as apresentadas no grupo de *steps Big Data*, que permitem trabalhar com bancos de dados não relacionais e técnicas de armazenamento e processamento distribuídos, podem ser aplicadas a fim de gerar um novo resultado.

REFERÊNCIAS

- AZEVEDO, Bernardo. **Processo de Publicação de Dados Abertos Multidimensionais em Banco de Dados NoSQL**. 2016. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – DCC/IM, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2016.
- BARBOSA, Thiago et al. **Applying Data Warehousing and Big Data Techniques to Analyze Internet Performance**. 2015.
- BERNERS-LEE, Tim. **Linked Data**. 2006. Disponível em: <<https://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 7 dez. 2018.
- BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. **The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities**. Scientific American, 2001.
- BIZER, C.; HEATH, T; BERNERS-LEE, T. **The Linked Data - The Story So Far**. Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
- COTTRELL, Les. **PingER Case Studies**. 2016. Disponível em: <<https://confluence.slac.stanford.edu/display/IEPM/PingER+Case+Studies>>. Acessado em: 7 dez. 2018.
- _____. **PingER and the Digital Divide**. 2011. Disponível em <<https://confluence.slac.stanford.edu/download/attachments/123309267/brochure.docx>>. Acesso em: 6 dez. 2018.
- EQUIPE GT. RNP. **LinkedDataBR: Exposição, Compartilhamento e Conexão de Recursos de Dados Abertos na Web (Linked Open Data)**. 2011. Documentação (Ciência da Computação) – Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2011. Disponível em: <https://memoria.rnp.br/pd/gts2010-2011/gt_linkeddatabr.html>. Acesso em: 8 dez. 2018.
- KIMBALL, Ralph; CASERTA, Joe. **The Datawarehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data**. Wiley Publishing Inc, 2004.
- NUNES, Raphaela. **Processamento Analítico Distribuído de Grandes Volumes de Dados Multidimensionais: uma Abordagem Baseada em Grafos RDF**. 2016. Dissertação (Mestrado em Sistemas em Informação) – Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2016.
- SLAC. **SLAC by the Numbers**. 2016. Disponível em: <https://www6.slac.stanford.edu/sites/www6.slac.stanford.edu/files/slac_by_the_numbers_facsheet_2016_final.pdf>. Acesso em: 6 dez. 2018.

_____. **SLAC Strategic Plan**. 2014. Disponível em:
<https://www6.slac.stanford.edu/files/Strategic_Plan_2014.pdf>. Acesso em: 6 dez. 2018.

SOUZA, Renan. **Processo de Publicação de Dados Abertos Interligados e Aplicação a Dados de Desempenho de Rede**. 2013. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – DCC/IM, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2013.

**APÊNDICE A – EXEMPLO DE ARQUIVO N-TRIPLES RDF: TRIPLAS
REFERENTES ÀS MEDIDAS DE THROUGHPUT NA GRANULARIDADE
HORÁRIA, PARA O DIA 01/07/2016³⁷**

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#Measurement> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasValue>
"269.627"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasSourceNode> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDestinationNode> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#www-05.nexus.ao> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDateTime> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Time2016Jul01H00> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#measuresMetric> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Throughput> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H00> <http://www-

³⁷ Neste apêndice são apresentadas as triplas correspondentes apenas às primeiras 6 medidas contidas no Anexo A.

iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasPacketSize> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#PacketSize100> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#Measurement> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasValue>
"269.591"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasSourceNode> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDestinationNode> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#www-05.nexus.ao> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDateTime> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#Time2016Jul01H01> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#measuresMetric> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#Throughput> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H01> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasPacketSize> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#PacketSize100> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://www-
iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#Measurement> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#hasValue>
 "269.620"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#hasSourceNode> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#hasDestinationNode> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#www-05.nexus.ao> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#hasDateTime> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Time2016Jul01H02> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#measuresMetric> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Throughput> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H02> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#hasPacketSize> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#PacketSize100> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H03> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#Measurement> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H03> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.ow#hasValue>
 "268.411"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H03> <http://www-

iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasSourceNode> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H03> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDestinationNode> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#www-05.nexus.ao> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H03> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDateTime> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#Time2016Jul01H03> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H03> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#measuresMetric> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#Throughput> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H03> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasPacketSize> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#PacketSize100> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H04> <http://www.w3.org/1999/02/22-rdf-
 syntax-ns#type> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#Measurement> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H04> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasValue>
 "269.438"^^<http://www.w3.org/2001/XMLSchema#float> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H04> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasSourceNode> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov> .
 <http://www- iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
 05.nexus.ao- Throughput-100-Time2016Jul01H04> <http://www-
 iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDestinationNode> <http://www-
 iepm.slac.stanford.edu/pinger/lod/resource#www-05.nexus.ao> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H04> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDateTime> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Time2016Jul01H04> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H04> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#measuresMetric> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Throughput> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H04> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasPacketSize> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#PacketSize100> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#Measurement> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasValue> "268.956"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasSourceNode> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDestinationNode> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#www-05.nexus.ao> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www-iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasDateTime> <http://www-iepm.slac.stanford.edu/pinger/lod/resource#Time2016Jul01H05> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www-

iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#measuresMetric> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#Throughput> .

<http://www-iepm.slac.stanford.edu/pinger/lod/resource#pinger-host.fnal.gov-www-
05.nexus.ao-Throughput-100-Time2016Jul01H05> <http://www-

iepm.slac.stanford.edu/pinger/lod/ontology/pinger.owl#hasPacketSize> <http://www-
iepm.slac.stanford.edu/pinger/lod/resource#PacketSize100> .

APÊNDICE B – DEMONSTRATIVO DOS PASSOS (“STEPS”) ELABORADOS PARA DADOS PINGER NA GRANULARIDADE DIÁRIA

Como forma de auxílio para o entendimento e a reprodução da solução sugerida neste trabalho, nas figuras listadas a seguir, são apresentadas todas as configurações realizadas em cada passo do fluxo elaborado no PDI para a solução SchedPingER em que os dados de entrada estejam na granularidade diária.

Passo 1 - *PingER Measurements TXT Input*

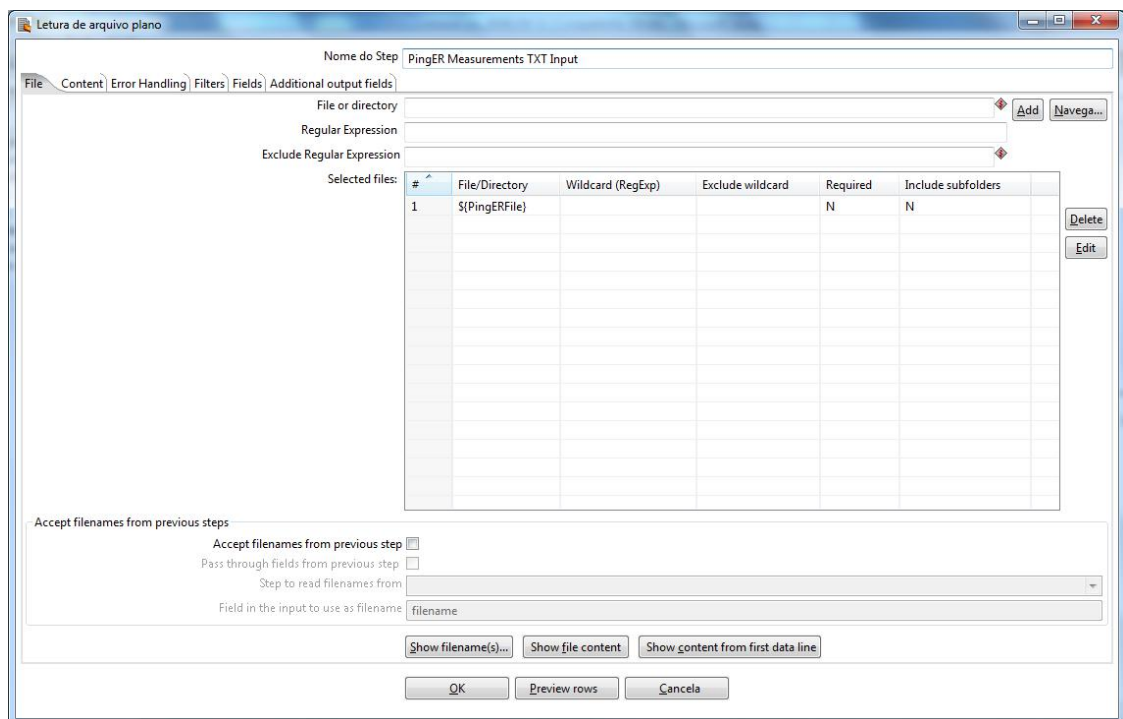


Figura 30 – Propriedades do Passo 1 da transformação para a granularidade diária, na aba “File”, indicando a variável que conterá o caminho para o arquivo de entrada da transformação

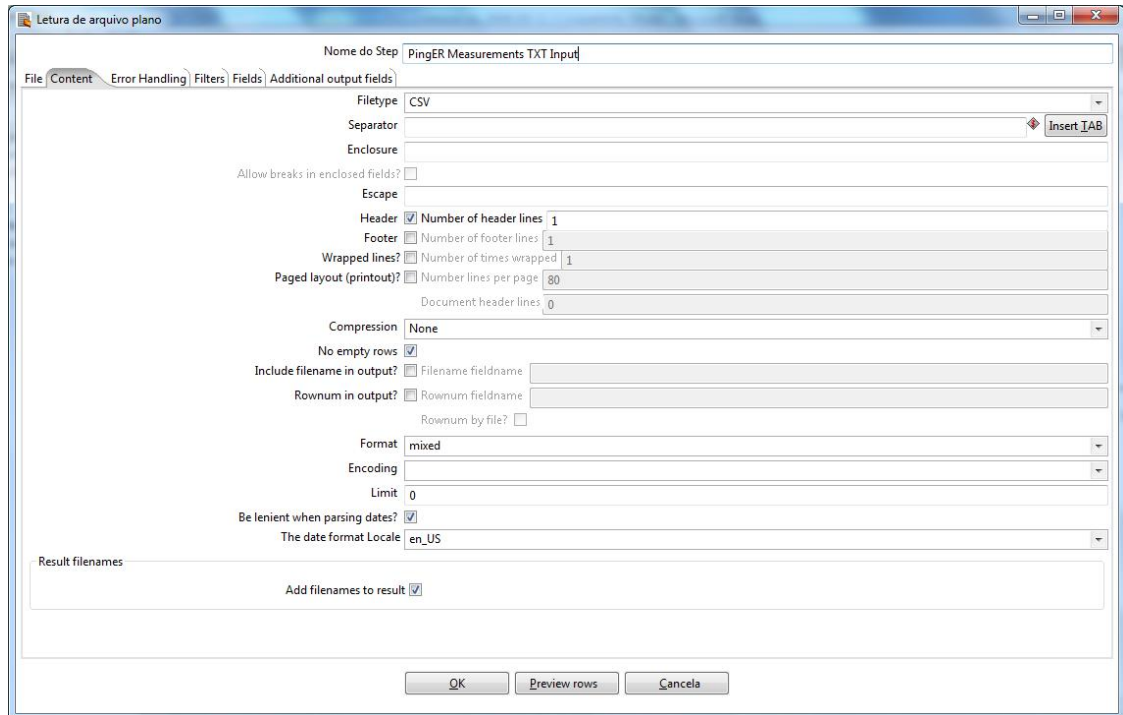


Figura 31 – Propriedades do Passo 1 da transformação para a granularidade diária, na aba “Content”, indicando características do conteúdo do arquivo, o seu formato e o caractere separador de colunas

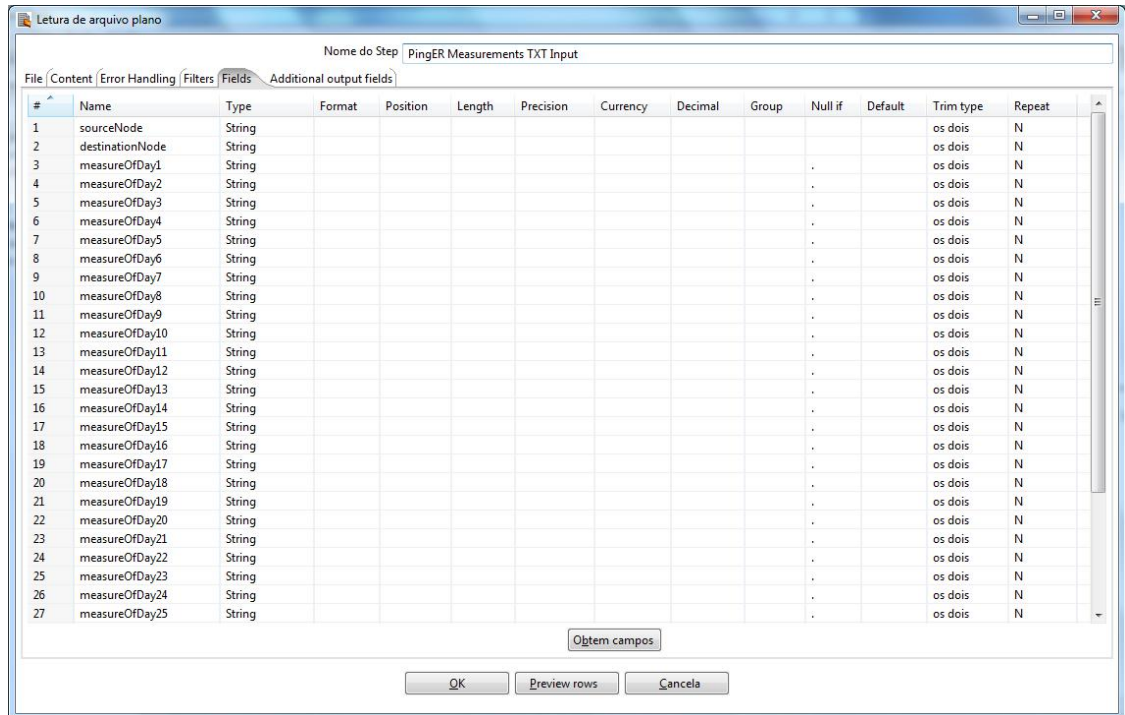


Figura 32 – Propriedades do Passo 1 da transformação para a granularidade diária, na aba “Fields”, em que são definidas cada coluna contida no arquivo de entrada, com a indicação do tipo do dado e do valor que representa o valor nulo

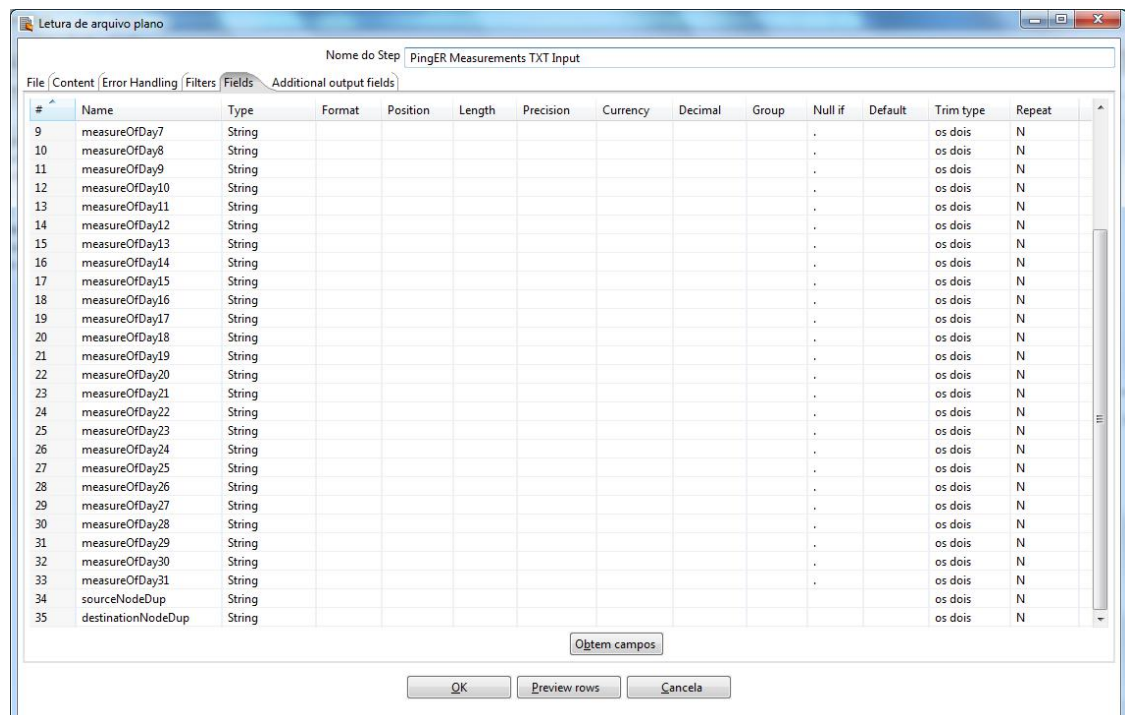


Figura 33 – Propriedades do Passo 1 da transformação para a granularidade diária, na aba “Fields”, em continuação às apresentadas na Figura 32

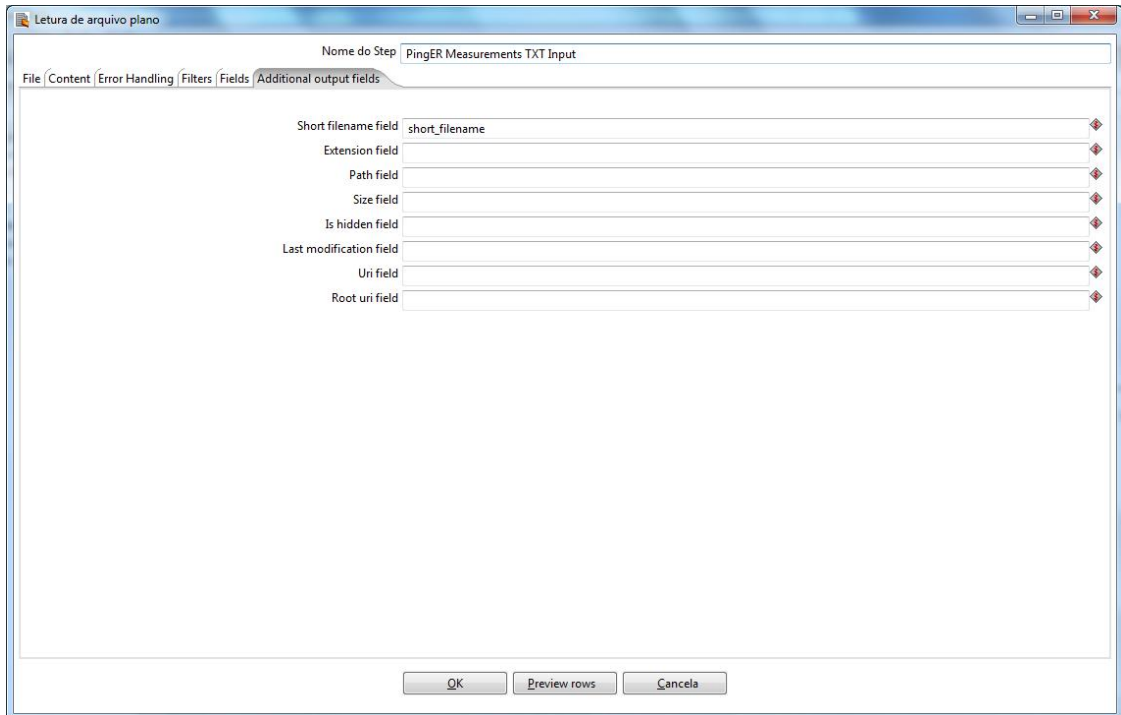


Figura 34 – Propriedades do Passo 1 da transformação para a granularidade diária, na aba “Additional output fields”, indicando que o nome do arquivo de entrada será disponibilizado para o passo seguinte

Passo 2 - Split Filenames Fields

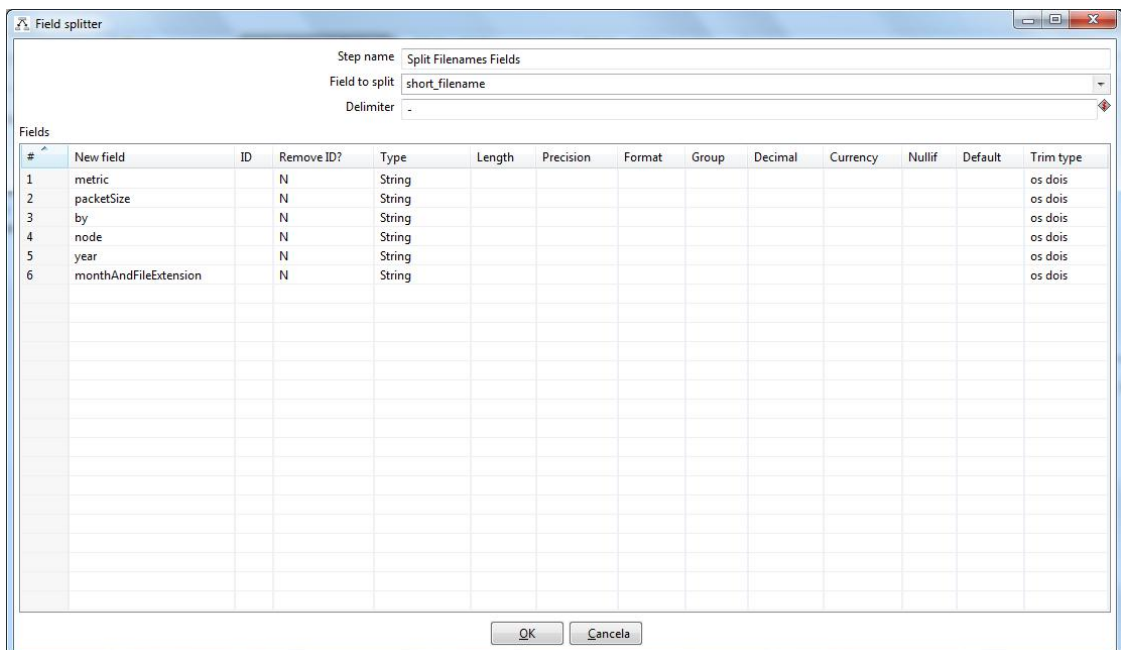


Figura 35 – Propriedades do Passo 2 da transformação para a granularidade diária, identificando cada informação contida no nome do arquivo de entrada

Value Mapper

Step name : Month Label Mapper

Fieldname to use : month

Target field name (empty=overwrite) : monthLabel

Default upon non-matching :

Field values:

#	Source value	Target value
1	01	Jan
2	02	Feb
3	03	Mar
4	04	Apr
5	05	May
6	06	Jun
7	07	Jul
8	08	Aug
9	09	Sep
10	10	Oct
11	11	Nov
12	12	Dec

OK Cancela

Figura 37 – Propriedades do Passo 4 da transformação para a granularidade diária, correlacionando cada número em referência a um mês do ano com a sua abreviatura correspondente

Passo 5 - *Metric Name Mapper*

Value Mapper

Step name : Metric Name Mapper

Fieldname to use : metric

Target field name (empty=overwrite) : metricName

Default upon non-matching : NOT-MATCH

Field values:

#	Source value	Target value
1	alpha	Directivity
2	average_rtt	AverageRTT
3	conditional_loss_probability	ConditionalLossProbability
4	duplicate_packets	DuplicatePackets
5	ipdv	IPDV
6	iqr	IQR
7	maximum_rtt	MaximumRTT
8	minimum_packet_loss	MinimumPacketLoss
9	minimum_rtt	MinimumRTT
10	MOS	MOS
11	out_of_order_packets	OutOfOrderPackets
12	packet_loss	PacketLoss
13	throughput	Throughput
14	unpredictability	Unpredictability
15	unreachability	Unreachability
16	zero_packet_loss_frequency	ZeroPacketLossFrequency

OK Cancela

Figura 38 – Propriedades do Passo 5 da transformação para a granularidade diária, a fim de padronizar cada nome de métrica considerada no projeto PingER

Passo 6 - *Script: Value Validation*

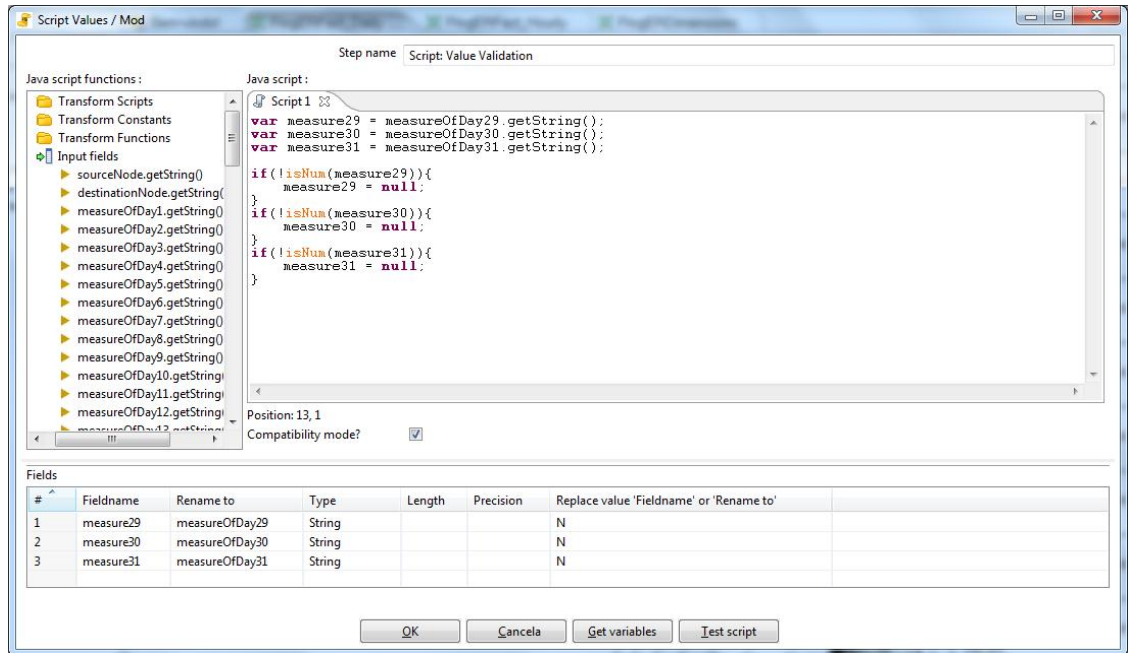


Figura 39 – Código contido no Passo 6 da transformação para a granularidade diária, para tratamento das variáveis que armazenam as medidas dos dias 29, 30 e 31: caso o valor seja numérico, retorna o valor medido; senão, retorna nulo

Passo 7 - Formula: URIs

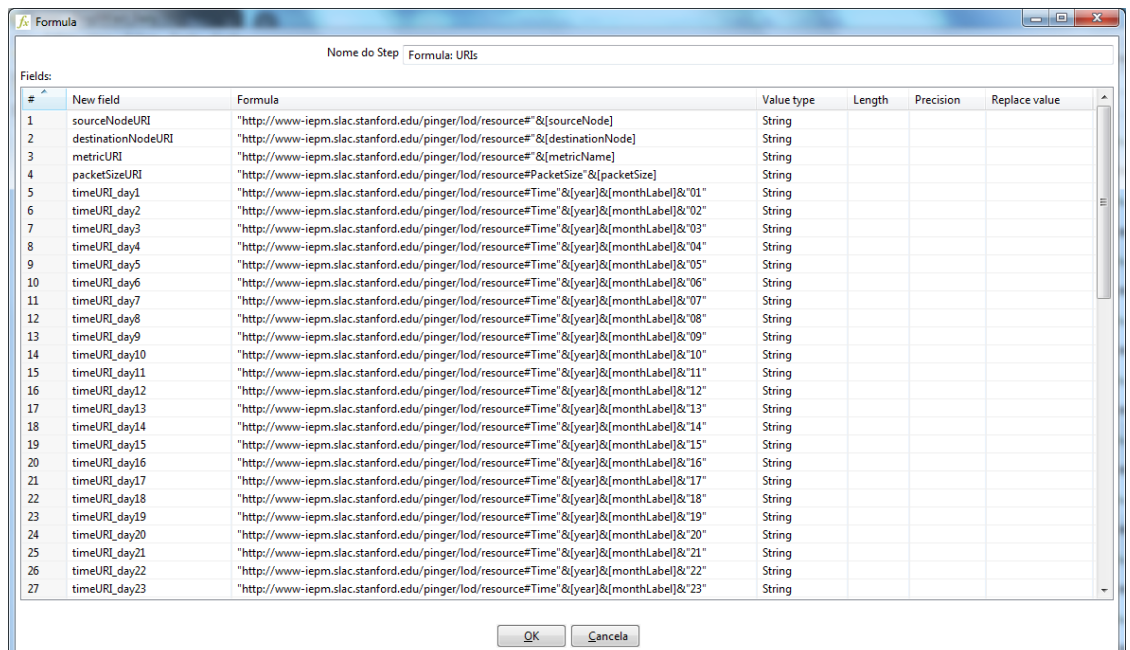


Figura 40 – Propriedades do Passo 7 da transformação para a granularidade diária, em que cada URI é descrito conforme a ontologia definida em (SOUZA, 2013)

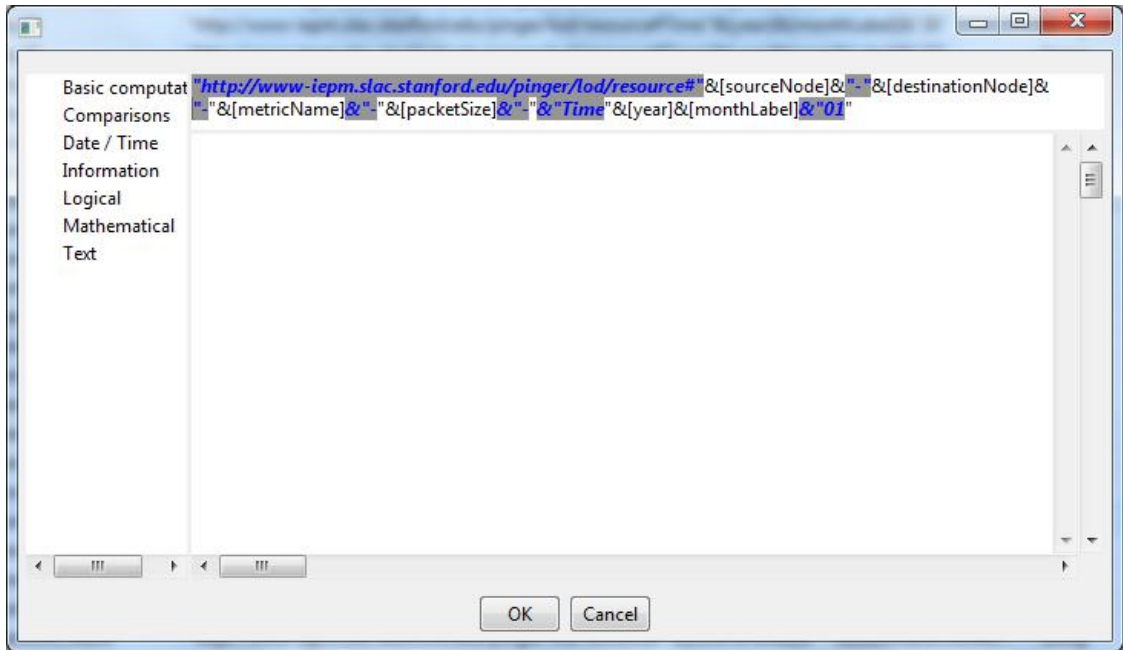


Figura 43 – Detalhamento da fórmula que representa cada URI da medida de um dia (*measurementURI_dayX*), conforme visto nas Figuras 41 e 42, onde X é um dia de 1 a 31 representado por um número de dois dígitos no final da definição do URI

Passo 8 - *Filter rows*

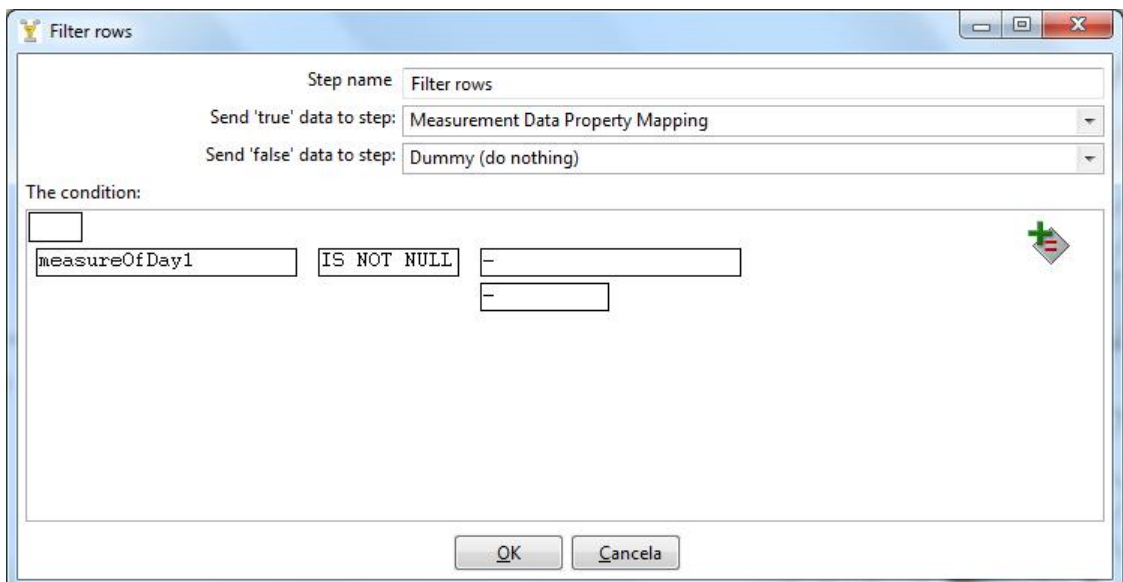


Figura 44 – Propriedades do Passo 8 da transformação para a granularidade diária, onde é verificada a existência de medida para o dia analisado

Passo 9 - *Measurement Data Property Mapping / Measurement Object Property*

Mapping

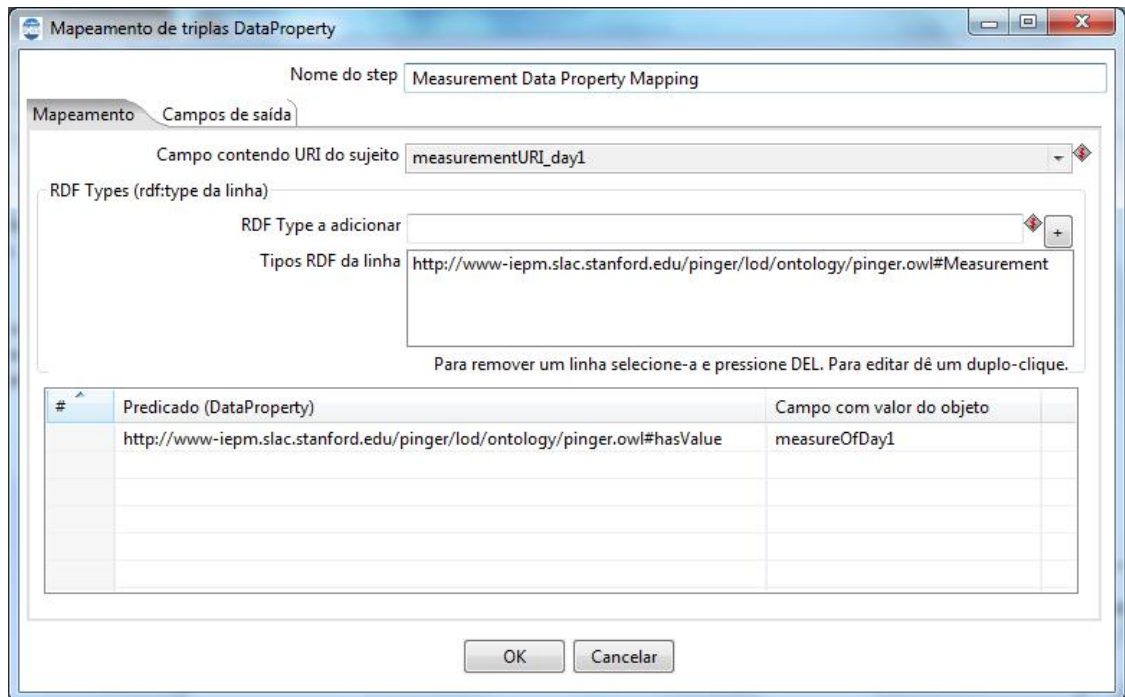


Figura 45 – Propriedades do Passo 9 da transformação para a granularidade diária, correspondente à aba “Mapeamento” do step “*Measurement Data Property Mapping*”, em que as triplas indicando que trata-se de uma medida e a de seu respectivo valor são criadas

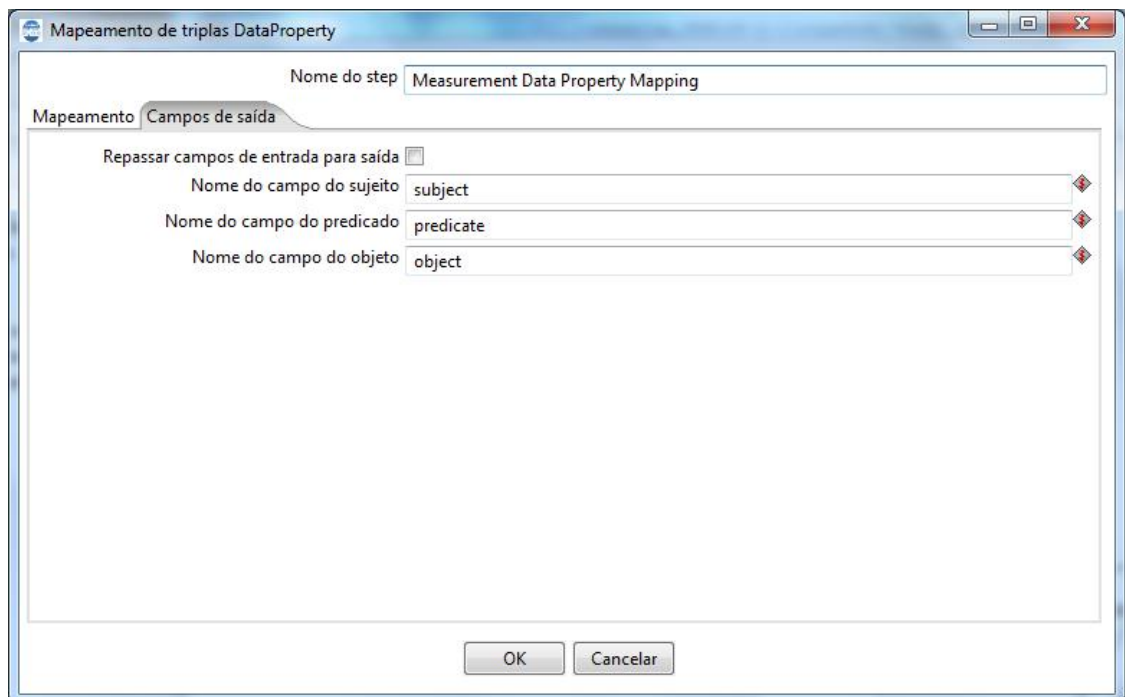


Figura 46 – Propriedades do Passo 9 da transformação para a granularidade diária, correspondente à aba “Campos de saída” do step “*Measurement Data Property Mapping*”, indicando como resultado os campos de sujeito, predicado e objeto que compõem a tripla

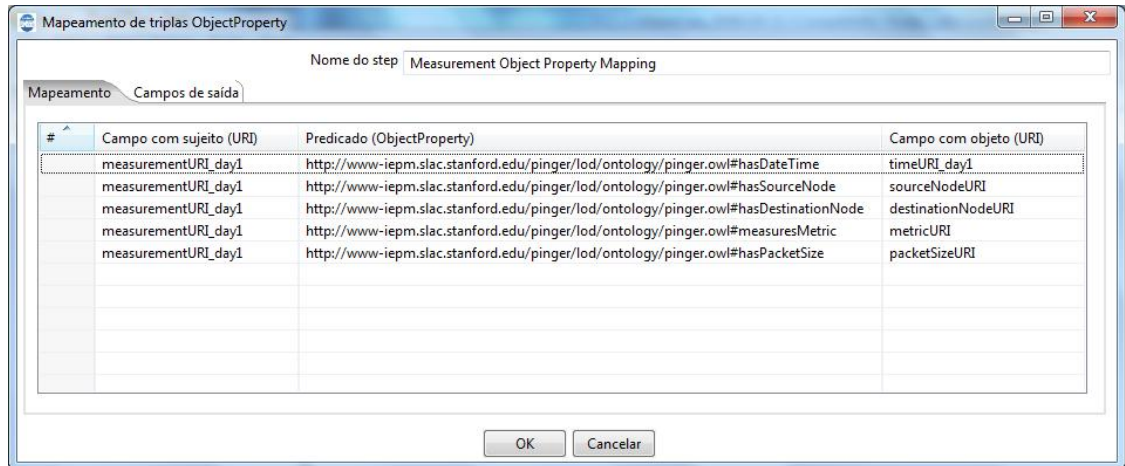


Figura 47 – Propriedades do Passo 9 da transformação para a granularidade diária, correspondente à aba “Mapeamento” do *step* “*Measurement Object Property Mapping*”, em que as triplas de cada parâmetro que define uma medida são criadas

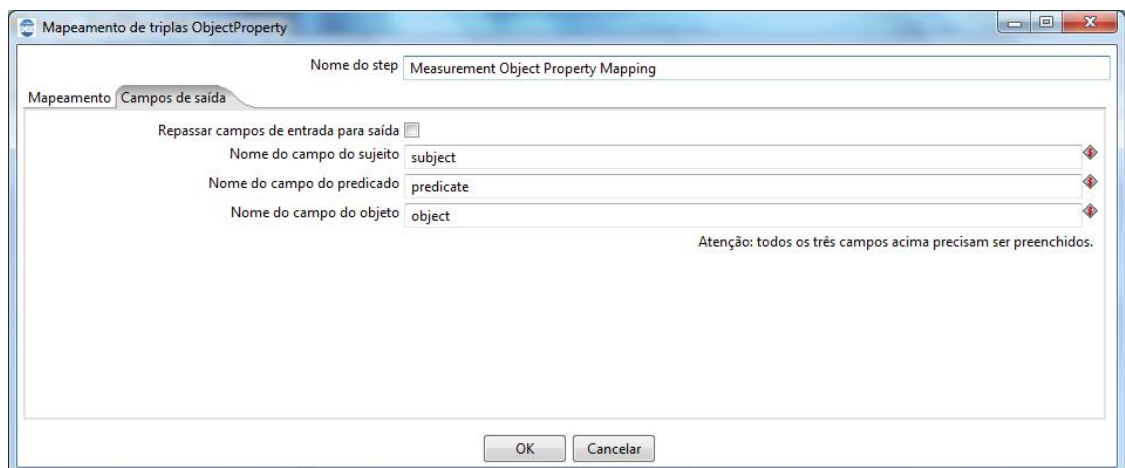


Figura 48 – Propriedades do Passo 9 da transformação para a granularidade diária, correspondente à aba “Campos de saída” do *step* “*Measurement Object Property Mapping*”, indicando como resultado os campos de sujeito, predicado e objeto que compõem a tripla

Passo 10 - *Add constants*

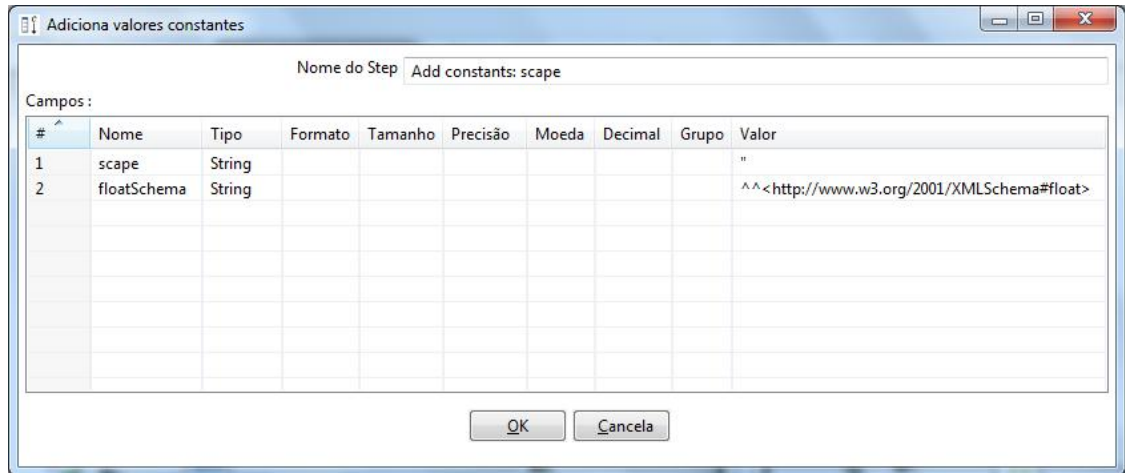


Figura 49 – Propriedades do Passo 10 da transformação para a granularidade diária, para definição das constantes necessárias na criação das triplas no formato N-Triples RDF

Passo 11 - *Formula: DataProp / Formula: ObjectProp*

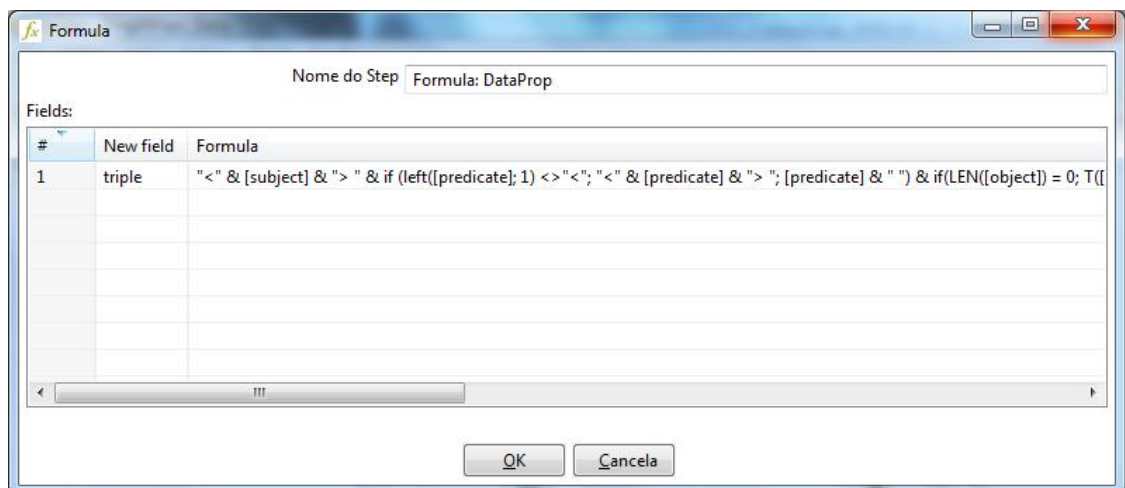


Figura 50 – Propriedades do Passo 11 da transformação para a granularidade diária, correspondente ao *step* “*Formula: DataProp*”, para a construção das triplas oriundas do *step* “*Measurement Data Property Mapping*”

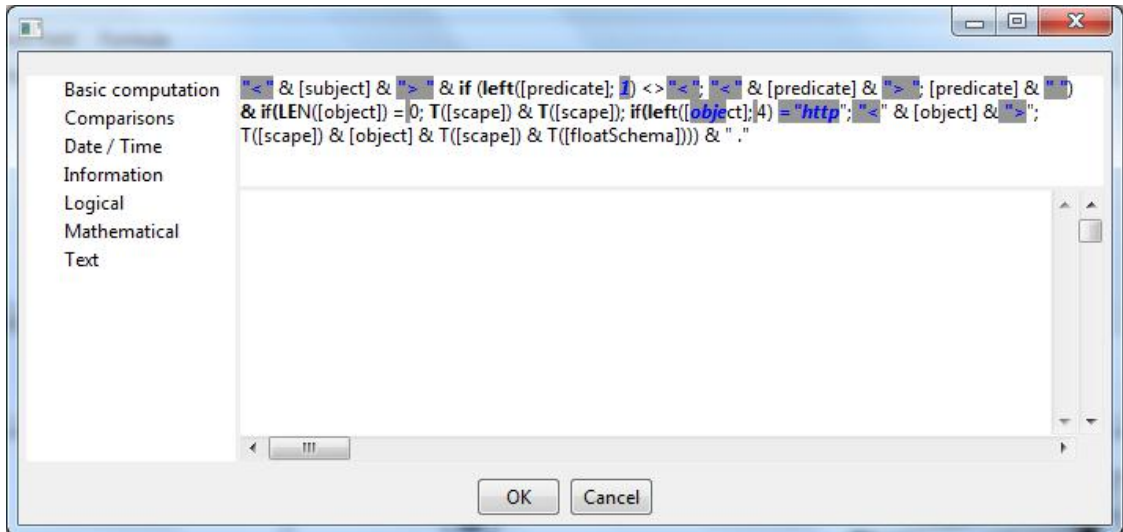


Figura 51 – Detalhamento da regra para a construção das triplas a partir dos campos de sujeito, predicado e objeto oriundos do *step* “*Measurement Data Property Mapping*”

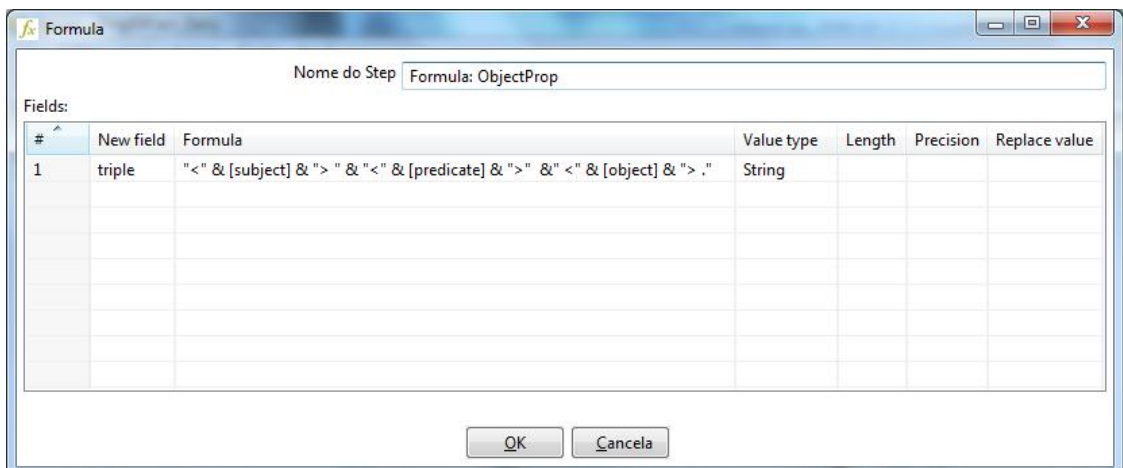


Figura 52 – Propriedades do Passo 11 da transformação para a granularidade diária, correspondente ao *step* “*Formula: ObjectProp*”, para a construção das triplas a partir dos campos de sujeito, predicado e objeto oriundos do *step* “*Measurement Object Property Mapping*”

Passo 12 - *DataProp Triples / ObjectProp Triples*

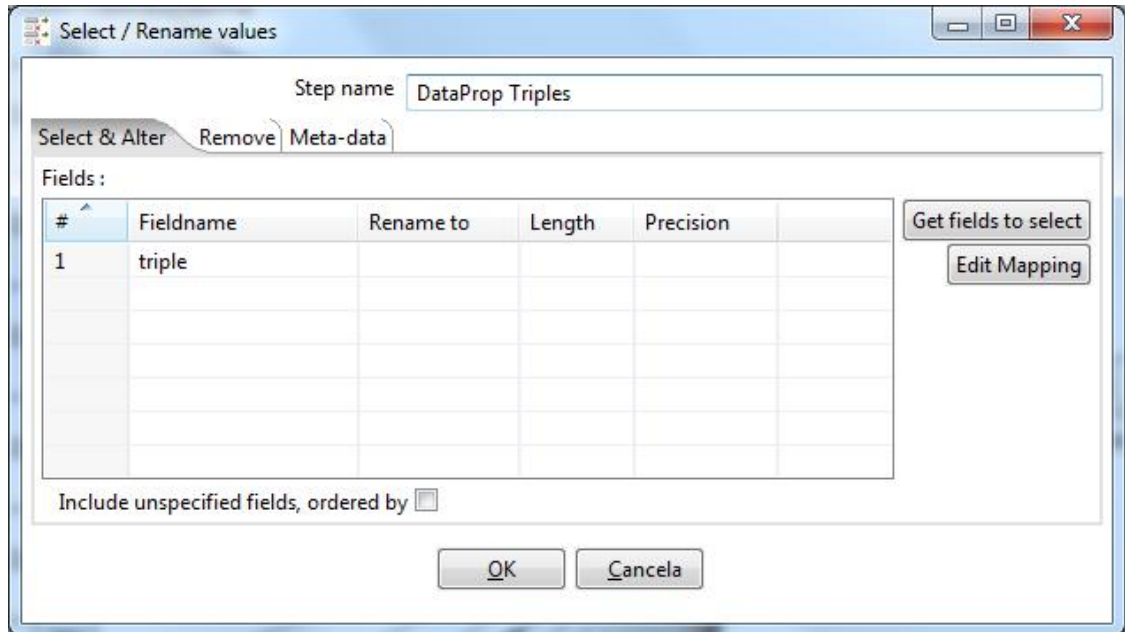


Figura 53 – Propriedades do Passo 12 da transformação para a granularidade diária, correspondente ao *step* “*DataProp Triples*”, em que a tripla construída no *step* “*Formula: DataProp*” é selecionada

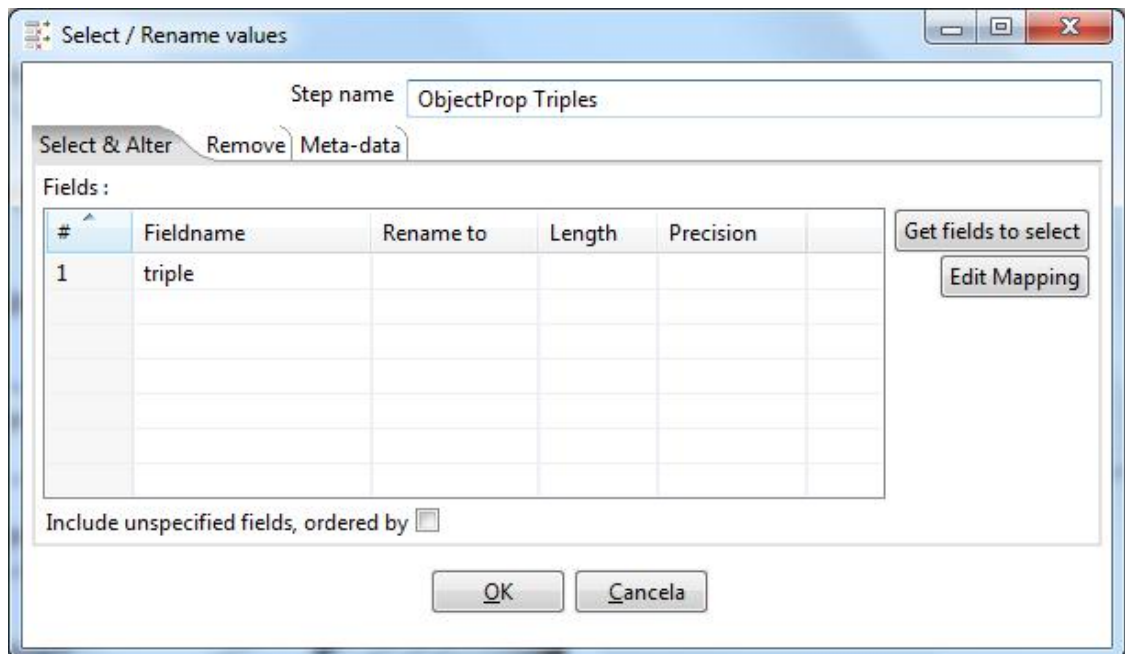


Figura 54 – Propriedades do Passo 12 da transformação para a granularidade diária, correspondente ao *step* “*ObjectProp Triples*”, em que a tripla construída no *step* “*Formula: ObjectProp*” é selecionada

Passo 13 - *Measurement Triples Output*

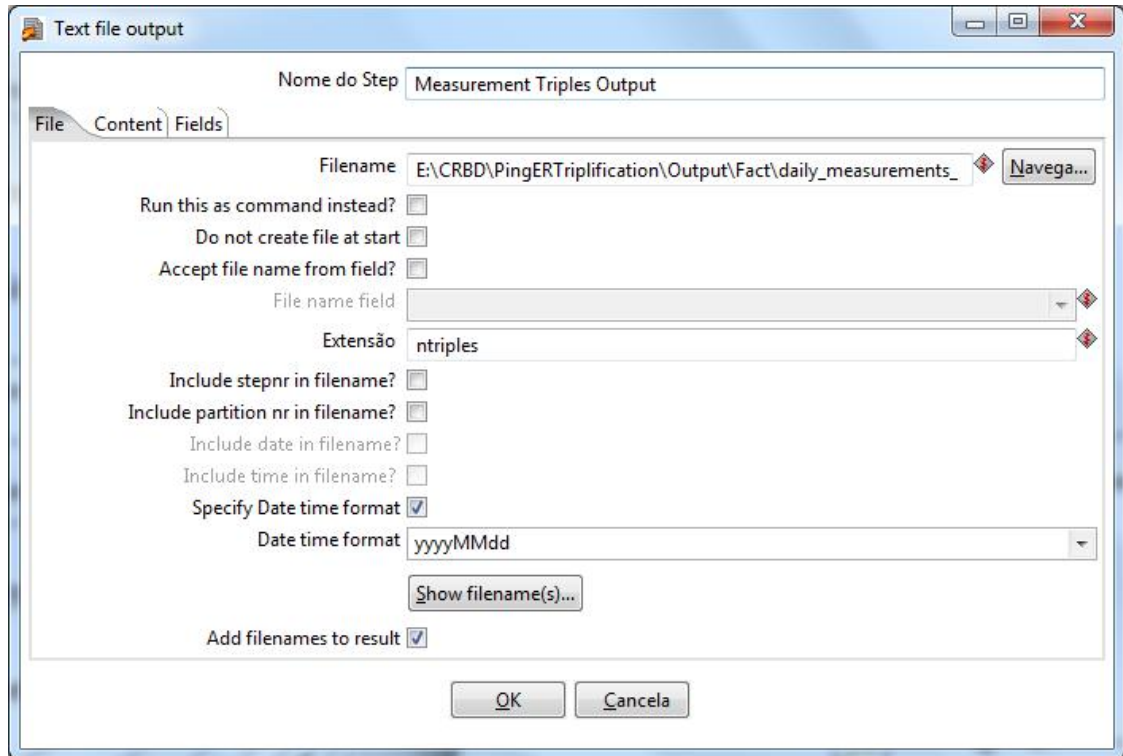


Figura 55 – Propriedades do Passo 13 da transformação para a granularidade diária, na aba “File”, em que são definidos o caminho e o formato do arquivo de saída

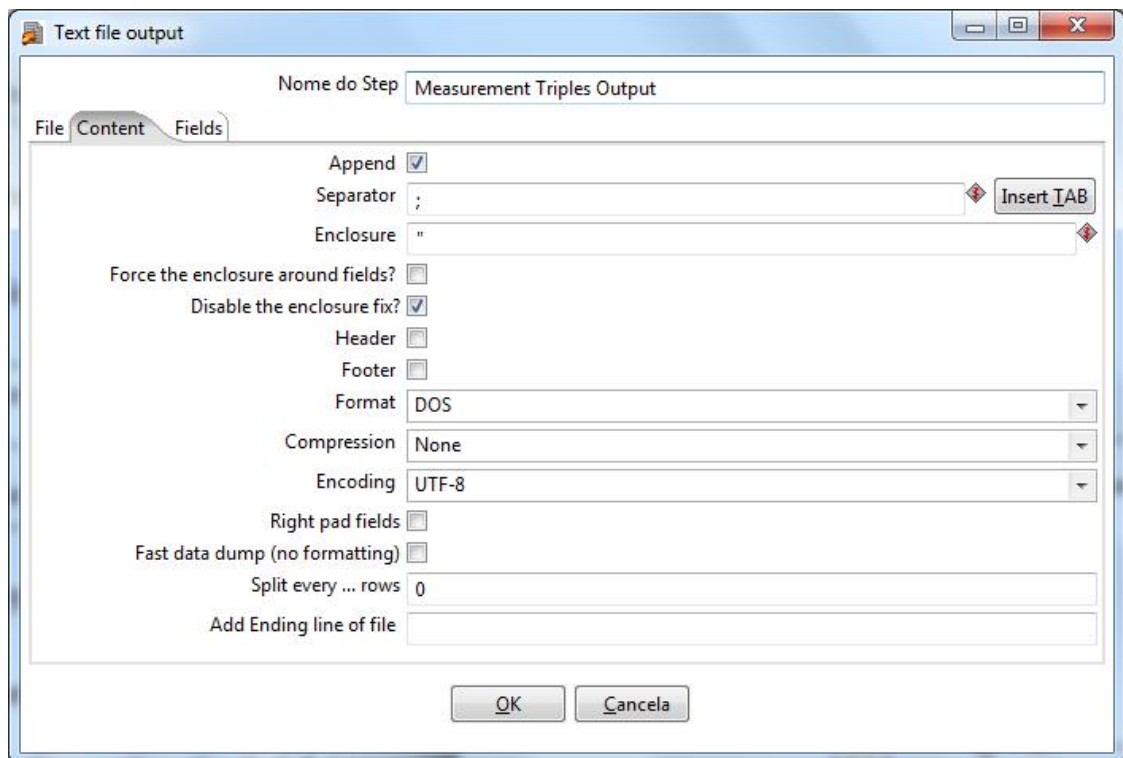


Figura 56 – Propriedades do Passo 13 da transformação para a granularidade diária, na aba “Content”, em que são definidas as possíveis características do conteúdo do arquivo

ANEXO A – EXEMPLO DE ARQUIVO PINGER: MEDIDAS DE THROUGHPUT NA GRANULARIDADE HORÁRIA, PARA O DIA 01/07/2016 (THROUGHPUT-100-BY-NODE-2016-07-01.TXT)³⁸

```

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
pinger-host.fnal.gov www-05.nexus.ao 269.627 269.591 269.620 268.411 269.438
268.956 269.457 269.136 269.431 269.005 10.204 267.313 269.485 269.452 269.179 269.365
269.373 269.425 269.561 269.553 269.582 269.592 269.424 268.846 ping-er-host.fnal.gov
www-05.nexus.ao
comsatsswl.seecs.edu.pk ping-er.usindh.edu.pk . . . . .
comsatsswl.seecs.edu.pk ping-er.usindh.edu.pk
ping-er.cern.ch speedtest.ivory.azstarnet.az 780.801 780.917 780.648 780.710 780.797
780.780 780.620 780.554 780.368 780.582 780.236 254.175 471.917 554.863 654.031
612.687 436.192 280.805 340.899 403.773 656.866 709.904 780.368 780.467 ping-er.cern.ch
speedtest.ivory.azstarnet.az
ping-er-raspberry.slac.stanford.edu ftp.physics.carleton.ca 1193.659 1194.065
1193.650 1193.254 1141.576 1150.000 1194.779 1193.457 1194.180 1193.361 1193.814
1189.766 1194.296 1194.026 1193.602 1194.827 1193.621 1195.649 1194.116 1194.065
1193.862 1193.071 1193.669 1194.142 ping-er-raspberry.slac.stanford.edu
ftp.physics.carleton.ca
ping-er.fsktm.um.edu.my speedtest.amnetdatos.com.ni 244.122 244.050 239.032
242.871 241.764 241.640 235.446 241.724 241.729 241.986 241.989 241.386 240.977
240.584 240.908 241.449 241.381 240.138 240.730 241.120 241.359 240.637 240.402
241.967 ping-er.fsktm.um.edu.my speedtest.amnetdatos.com.ni
ping-er.stanford.edu ping.rmki.kfki.hu 430.406 430.336 430.231 430.325 430.236
430.476 430.609 430.480 430.386 430.521 430.575 430.562 430.545 430.429 430.558
430.515 430.378 430.475 430.448 430.341 430.371 430.264 430.297 430.315
ping-er.stanford.edu ping.rmki.kfki.hu
ping.riken.jp www.hep.uiuc.edu . . . . . ping.riken.jp
www.hep.uiuc.edu
pingersonar-um.myren.net.my www.umt.edu.my 4432.027 4403.756 4394.848
4398.250 4391.843 4418.771 4402.837 4431.495 4399.429 4420.490 4420.225 4382.594

```

³⁸ Neste anexo é apresentada apenas uma fração do arquivo original (45 linhas do total de 6180 linhas.)

4456.088 4409.408 4421.151 4401.919 4403.362 4392.366 4421.945 4430.963 4405.857
4415.602 4401.657 4398.119 pingersonar-um.myren.net.my www.umt.edu.my
perfonar-unimas.myren.net.my ps-lt.aura.ampath.net 170.828 170.642 168.951
170.736 170.817 169.882 165.035 164.411 170.970 170.961 170.886 170.876 168.495
170.947 170.909 170.828 166.646 166.604 166.663 166.665 165.872 166.616 168.759
126.118 perfonar-unimas.myren.net.my ps-lt.aura.ampath.net
netmon.physics.carleton.ca ftp.physics.carleton.ca
netmon.physics.carleton.ca ftp.physics.carleton.ca
pinger-host.fnal.gov www.ml.refer.org 371.237 367.359 357.097 370.583 371.111
372.856 373.051 372.556 373.032 370.662 370.939 361.784 369.796 330.944 372.638
368.103 372.895 371.577 372.913 371.982 372.664 372.517 360.620 370.604 pinger-
host.fnal.gov www.ml.refer.org
pinger.slac.stanford.edu www.hepi.edu.ge 340.163 340.063 340.239 340.197 340.309
340.204 332.425 340.128 337.575 337.084 333.487 319.894 330.918 340.128 340.182
340.229 340.153 339.916 340.174 339.954 338.439 340.154 340.272 340.186
pinger.slac.stanford.edu www.hepi.edu.ge
netmon.physics.carleton.ca www.puc.cl 457.404 470.515 458.825 464.596 469.021
467.537 460.254 466.062 469.021 470.515 470.515 467.537 464.596 328.519 450.432
464.596 469.021 447.702 466.062 464.596 467.537 467.537 466.062 466.062
netmon.physics.carleton.ca www.puc.cl
pingersonar-utm.myren.net.my www.polmed.ac.id
pingersonar-utm.myren.net.my www.polmed.ac.id
pinger.slac.stanford.edu www.ubd.edu.bn 372.980 373.193 371.672 370.980 366.815
366.348 367.752 366.348 365.821 269.017 362.614 197.808 187.152 264.320 362.320
263.824 186.532 263.868 166.073 359.868 365.139 367.882 369.071 364.168
pinger.slac.stanford.edu www.ubd.edu.bn
ping.riken.jp www.in2p3.fr ping.riken.jp www.in2p3.fr
pinger.aiou.edu.pk www.mtec.by 352.980 358.602 359.372 364.452 359.343 355.398
358.660 325.083 364.343 358.212 357.997 359.421 363.089 359.149 355.096 359.254
355.448 354.373 353.435 358.908 358.947 359.722 359.975 359.723 pinger.aiou.edu.pk
www.mtec.by
pinger.slac.stanford.edu utdallas.edu 919.523 920.010 920.916 920.652 922.037
920.474 921.553 920.560 921.766 919.580 921.335 920.422 921.191 920.738 921.737

920.010 907.381 800.655 796.167 733.475 810.707 896.126 751.567 742.984
 pinger.slac.stanford.edu utdallas.edu

pinger.unesp.br mail.lttnet.net 270.370 261.770 272.582 273.502 273.087 271.504
 269.344 269.386 269.831 269.663 269.663 270.148 269.719 269.437 269.260 269.350
 269.744 269.173 269.852 269.465 269.322 269.446 269.826 269.395 pinger.unesp.br
 mail.lttnet.net

iepm01.desy.de www.daffodilvarsity.edu.bd
 iepm01.desy.de www.daffodilvarsity.edu.bd

sitka.triumf.ca www.suvl.edu.sd sitka.triumf.ca
 www.suvl.edu.sd

pinger.slac.stanford.edu www.unikl.edu.my 394.236 394.102 394.316 393.318
 394.116 393.987 393.128 394.157 393.997 394.173 394.021 393.999 394.246 394.089
 279.850 393.961 394.106 394.071 394.388 394.452 394.277 394.447 394.397 394.532
 pinger.slac.stanford.edu www.unikl.edu.my

pinger.nwfpuet.edu.pk win17.b2web.com.ve 191.988 259.188 115.468 139.177
 119.768 188.067 192.124 90.333 259.305 140.036 253.491 76.521 136.360 73.796 67.000
 78.869 72.806 90.339 135.185 135.120 75.979 107.586 257.171 244.284
 pinger.nwfpuet.edu.pk win17.b2web.com.ve

pinger.daffodilvarsity.edu.bd www.lsx.com.la 441.662 205.232 336.265 465.502
 344.990 465.480 464.905 456.607 438.749 400.804 449.686 292.528 387.987 401.472
 311.438 419.942 441.584 452.259 453.664 454.356 439.965 427.648 427.964 432.305
 pinger.daffodilvarsity.edu.bd www.lsx.com.la

pinger.uum.edu.my www.myren.net.my 5502.071 5517.482 5531.322 5521.606
 5502.686 4083.791 4997.518 5499.204 5493.070 3969.521 5522.225 5511.307 5497.158
 5516.040 5505.967 5532.772 5510.074 5522.019 5512.335 5522.019 5510.279 5542.943
 5515.834 5518.925 pinger.uum.edu.my www.myren.net.my

pingeramity.in www.sierratel.sl 234.539 234.408 234.842 242.379 235.510 235.347
 231.326 232.545 232.925 231.970 232.850 234.506 233.065 234.404 233.908 232.234
 233.184 231.895 233.007 232.474 232.804 232.304 231.589 231.488 pingeramity.in
 www.sierratel.sl

pinger.nwfpuet.edu.pk duhs.seecs.edu.pk
 pinger.nwfpuet.edu.pk duhs.seecs.edu.pk

pingersonar-um.myren.net.my www.ub.edu.ph 366.479 363.523 366.274 369.580
 363.233 359.387 344.076 255.527 354.325 367.247 363.364 369.922 360.209 367.562

351.468 350.543 351.042 350.134 360.457 369.998 367.384 364.429 367.086 363.488
pingersonar-um.myren.net.my www.ub.edu.ph

 pinger.slac.stanford.edu www.irk.ru 287.862 287.924 287.932 287.915 287.681
287.830 287.630 287.628 287.684 287.712 287.758 287.828 287.865 287.788 287.636
287.769 287.832 287.756 287.859 287.782 287.834 287.990 287.920 287.908
pinger.slac.stanford.edu www.irk.ru

 pinger.uum.edu.my www.it.su.se 327.829 327.857 327.753 327.834 327.816 327.096
327.662 327.854 327.708 327.822 313.994 314.028 313.834 314.051 313.842 314.010
313.996 314.048 314.040 314.004 313.994 314.077 313.997 314.017 pinger.uum.edu.my
www.it.su.se

 ping.riken.jp www.uem.mz 224.678 224.684 224.724 224.687 224.698 224.712
224.674 224.591 224.331 224.350 224.449 224.512 224.541 224.500 224.618 224.588
224.634 224.670 224.627 224.673 224.656 224.668 224.669 224.674 ping.riken.jp
www.uem.mz

 pinger.aiou.edu.pk www.esmt.sn 306.126 307.769 307.726 309.232 299.094 286.045
186.473 259.936 264.177 227.916 286.933 300.049 304.380 307.432 307.963 283.022
306.037 306.374 306.067 307.680 307.547 306.052 307.889 306.104 pinger.aiou.edu.pk
www.esmt.sn

 rainbow.inp.nsk.su dns1.ethz.ch rainbow.inp.nsk.su
dns1.ethz.ch

 pinger.nchc.org.tw www.ait.ac.th 448.644 448.403 423.736 398.083 421.919 450.145
450.004 449.702 449.806 446.756 446.152 449.838 446.768 447.404 449.183 446.487
447.759 450.171 448.840 450.381 447.573 449.607 449.869 449.138 pinger.nchc.org.tw
www.ait.ac.th

 pinger.uum.edu.my www.mimos.my 6087.417 2763.238 5865.090 5742.445 6090.428
5976.602 5884.948 4270.606 5052.723 6047.796 6216.511 6328.619 5869.750 6321.038
6307.544 6227.779 6145.658 6198.775 6011.377 6131.884 5791.064 5987.987 6309.969
6321.849 pinger.uum.edu.my www.mimos.my

 121.52.146.180 dxcnaf.cnaf.infn.it 220.065 220.048 220.060 220.027 220.024
220.064 220.063 220.025 220.058 220.012 220.059 220.032 220.032 220.020 219.817
219.610 219.525 219.376 219.276 219.524 219.589 219.613 219.601 219.583 121.52.146.180
dxcnaf.cnaf.infn.it

ping.riken.jp mail.gnet.tn 260.128 259.757 259.821 259.608 259.929 259.726 259.813
 259.609 259.671 259.763 259.863 259.642 259.685 259.479 257.285 259.715 259.367
 259.739 259.744 259.776 259.620 259.697 259.654 259.620 ping.riken.jp mail.gnet.tn

pinger.ascr.doe.gov www.lsx.com.la 241.840 141.391 245.931 231.154 258.407
 258.393 236.833 190.568 258.163 258.233 258.520 258.602 258.775 258.754 257.738
 241.210 258.230 258.436 258.736 258.901 258.401 258.896 258.736 258.190
 pinger.ascr.doe.gov www.lsx.com.la

pinger.fsktm.um.edu.my cdcnet.uniandes.edu.co
 pinger.fsktm.um.edu.my cdcnet.uniandes.edu.co

pinger.slac.stanford.edu 193.222.119.6 463.291 465.411 466.031 464.718 465.861
 466.035 465.536 463.297 465.868 465.824 465.621 465.931 465.441 465.523 465.890
 465.158 465.648 464.032 465.938 465.947 465.659 465.874 465.921 465.140
 pinger.slac.stanford.edu 193.222.119.6

pinger.cern.ch mail.gnet.tn 1198.171 1108.032 1140.932 1161.372 1161.372 1161.865
 1161.098 1161.692 1160.998 1160.861 1162.002 1160.907 1160.569 1161.573 1161.957
 1161.399 1160.350 1161.317 1160.961 1161.244 1161.710 1161.783 1161.792 1160.925
 pinger.cern.ch mail.gnet.tn

pinger.unesp.br ping.cern.ch 265.010 265.069 264.997 276.411 265.041 265.114
 265.060 265.065 265.111 264.979 264.979 135.447 264.884 264.893 264.930 264.912
 264.520 264.761 264.864 264.703 265.005 265.006 265.062 264.843 pinger.unesp.br
 ping.cern.ch

perfsnar.myren.net.my www.usep.edu.ph 909.386 912.042 897.471 856.612 879.352
 507.119 721.676 641.329 696.927 776.905 866.896 288.239 162.492 239.106 205.071
 161.879 205.080 891.261 913.101 903.143 910.395 905.291 905.651 903.331
 perfsnar.myren.net.my www.usep.edu.ph

multivac.sdsc.edu kadri.ut.ee 392.930 396.090 385.748 393.978 326.862 396.090
 393.978 343.585 395.031 395.031 393.978 396.090 376.892 397.155 371.210 341.205
 397.155 396.090 397.155 397.155 397.155 397.155 396.090 397.155 multivac.sdsc.edu
 kadri.ut.ee