Datagenerator

version 0.30

by: uwe geercken
web: www.datamelt.com
email: uwe.geercken@datamelt.com

last update: 2011-06-22

Table of Contents

Functional description

General

All programs that are created, need to be tested for completeness, functionality and quality. Most programs process data in one form or the other and so data is needed to check, if the program does what is was designed for.

Preferably there is data available from e.g. a source system that provides “real-life” data. Or there is a test system around, from which data can be taken to test. But sometimes that is either not the case, the amount of data is insufficient or the number of combinations within the data is too small to make up a representable test.

This is the scenario for the datagenerator tool. It generates mass data based on word lists, purely random data, datetime values or based on a regular expression pattern. The data is output to an ASCII file with fields of fixed length or field divided using a separator. From there it can be used for testing of programs or for feeding it into a tool that processes the data. The output format of the data is specified in a simple XML file, defining the content and the length of the fields which make up the rows in the output file.

How it works

Word lists provide a means to group similar items together such as e.g. “colors” or “weekdays” or “top 100 used english words”, just to name some. The tool randomly selects values from such a word list and uses the retrieved value for a specified field in the output row. As theses lists are text files – containing one value per row – they can easily be extended if necessary. You can – as an example - also create similar lists of words but in different languages. The advantage of word lists is, that you get real data that makes sense in your output file.

Another method of generating data, is to do it purely randomly. This is the easiest way of generating data. All that needs to be specified is the length of the data that shall be generated and the program starts to randomly generate text strings. Once the data is generated, you can e.g. mass load it into a database and test a program for performance or you could see if the tool can cope with an input file of 10 million lines. Of course the random data generated is senseless data. It has no meaning in real life.

There is a special type “datetime” that you can use to generate date and/or time values based on a given pattern. The format of the pattern follows the formatting of the pattern of the Java SimpleDateFormat class.

If we would generate e.g. multiple datetime fields and with a different pattern then we would get two independent dates – they are not related to each other. That is probably not what we want. We want different date fields which are correctly related to each other: So maybe we want a date field showing “2011-01-01” and another field with the weekday name equal to “Saturday” and not a randomly picked weekday. The datagenerator allows you to do so by defining reference fields using an id and then later referencing this id from other datetime fields. But also the other types such as regular expression, random and categories may be used as reference fields.

Furthermore the datagenerator allows you to construct (concatenate) a field from multiple reference fields, This can be used to combine the results of multiple fields into one field and thus gives the user more flexibility on the output.

The last way to generate data, is to use regular expression patterns. If you know regular expressions – which are quite common in many areas and across different operating systems – then you know that you can use them to test if data fits to a specified pattern. E.g. you have an email address and want to know if it is properly formatted. If you only have a few, you can do it in your head, but if you need to do it programmatically – e.g. on a web page - or if you have large amounts of email addresses, then regular expressions are a solution to do so: you have the data and test it against a pattern. The datagenerator program uses regular expression patterns exactly the other way around: you specify a pattern data should fit to and the tool generates masses of data according to it.

Output format

Below is an example of the XML file used to define how the output of the generated data should be formatted.

Please note that the reference fields are introduced in this version 0.30 – older versions do not have this feature.

```
<xml>
        <references>
                <field type="datetime" id="date1" pattern="yyyy-MM-dd"/>
                <field type="datetime" id="year" reference="date1" pattern="yyyy"/>
                <field type="category" id="airport" category="airport_code"/>
        </references>
        <row type="delimited" seperator=";">
                <field type="category" category="seasons" length="20" />
                <field type="datetime" pattern="q" length="1" />
                <field type="random" length="10" />
                <field type="regex" pattern="www.[a-pA-P]{8,16}.com" length="30" />
                <field type="reference" reference=”date1” length="10" />
                <field type="reference" reference=”airport/-/year” length="20" />
        </row>
</xml>
```

The example that comes with the program, contains information of how the xml file should be structured. If you want to (optionally) specify reference fields – fields that can be re-used by other fields – then you need to specify the fields as individual lines within the “references” tags. Each reference field needs to have an unique “id” attribute, which is used for referencing this field. Additionally fields of type “random” or “regex” need to have a “length” attribute defined and fields of type category need the “category” attribute specified.

A reference field itself can also reference another reference field. This is specifically helpful, when using datetime type fields. By referencing a datetime type field, the same datetime is used as the basis for the field referencing the other, resulting in date fields that correlate to each other. E.g. if you specify a datetime field “A” with a pattern of “yyyy-MM-dd” then the generated date might be e.g. “2011-06-21”. A datetime field “B” referencing this, will be based on the same date, So if you specify a pattern of “q” (the quarter of a date) then the generated value would be “2” because the date value generated for field “A” is in the second quarter of the relevant year.

Reference fields themselves are not output to the resulting target file. They are only used as a reference by defining a field of type “reference” and then using the “id” of the referenced field in the attribute “reference” of the field tag.

After the optional definition of the reference fields, specify the “row” tag and its type. Possible values are “fixed” and “delimited”. If you choose “fixed” then all fields and thus all rows will have the same length: if the generated data is shorter than the specified length of the field – for example if a short word is taken from a word list – then the remaining space, up to the length of the field, is filled with spaces. On the other hand, if you specify “delimited” as row type, then you also need to specify the “separator” attribute, which defines the separator to be used between the individual fields.

Inside the “row” tag, you can have as many field tags as you require to make up a full row of output data. Fields are of three different types: for taking data from a word list (type=category), for generating random data (type=random) or for generating data based on a regular expression pattern (type=regex). For details see above.

Independent of the type you define, you always need to specify the “length” attribute of the field, defining what the maximum length of a field should be.

If you choose “category” as the type of a field, then you also need to specify the name of the file to choose in the attribute “category”. So if your category file on the filesystem is named “colors.category”, then you should specify: category=”colors”. The values for the field will thus be randomly taken from this file. Category files always need to have the extension “.category”.

If you choose “random” as field type, then all you need to specify is the attribute “length” and data will randomly be generated up to this length.

Lastly, if you want to get data that is according to a regular expression pattern, then you need to define the “pattern” attribute. It shall contain a valid regular expression – in terms of Java regular expressions. E.g. the regular expression pattern “[a-zA-z0-9]” defines to use any lowercase alphabetic characters, any uppercase alphabetic characters or a number to be used. Another example would be to generate data according to following pattern: “[s-zäüö]{8,16}”. This would generate data according from the range of characters (s to z lowercase, plus the special characters ä, ü or ö). And it would generate strings of data from length 8 up to length 16. So you would get data of e.g. length 8, 12, 13, etc. If you leave away the second value in the brackets, then you define a fixed length of the data to be generated. So defining the pattern: “[s-zäüö]{8}” will always generate data of length eight (8).

But there are some limitations: In regular expressions you can define the characters “.”, “*” or “?” as placeholders for “any character”. The difference is, that they define the occurrence of none, one or many times of a character or group of characters. This is not possible with the datagenerator tool, as the tool would not know, how many characters to generate. Also – at the time being an “or” character (|) can not be used.

Please note the last two lines. It shows how to reference fields. In the “reference” attribute specify which field to reference by listing its “id” value. You can also use multiple reference fields by deviding each field with a slash character (“/”).

Example: reference=”reference1/-/reference2/:/reference3”

Here we use three reference fields for the definition of an output field. When a reference can not be found the literal value will be used instead. So in the example above the “-” and the “:” definition are output like this. The result of the example above could be:

Example output: value1-value2:value3

Using reference fields you have the liberty to create multiple output columns of type datetime that are based on the same date value and also to concatenate reference fields into one value.

Program properties file

There are several parameters that can be passed to the datagenerator program. Alternatively, all parameters can be defined in the properties file “datagenerator.properties”. In this case, no parameter shall be passed to the program, as it takes all arguments from the properties file.

Below find the parameters that can be specified in the properties file:

Name	*Description*	*Example*
numberofoutputlines	optional.specifies how many output lines should be generated	numberofoutputlines=100000
outputinterval	optional. specifies how often the number of generated rows will be displayed	outputinterval=1000
categoryfilesfolder	required. path to the folder where the category files are located	categoryfilesfolder=/home/generator/categories
rowlayoutfile	required. name and path of the file, describing the output format	rowlayoutfile=/home/generator/rowlayout.xml
outputfile	optional. name and path of the file which contains the generated data	outputfile=/home/generator/output.txt
verbose	optional. output status during processing	verbose=true
format	optional. specifies the output format: 0=regular (mixed) case, 1=lowercase only or 2=uppercase only	format=2
possiblevalues	optional. defines which character set should be used when generating random data	possiblevalues=ABCDEFGHIJKLMNOPQRST1234567890!$+*%&
maximumyear	optional. Defines the maximum year used during date generation. Default is 2199	maximumyear=2013

minimumyear	Optional. Defines the minimum year used during date generation. Default is 1970	minimumyear=2006

Program arguments

If you do not want to use a properties file, then you can pass all required arguments directly to the program.

Name	*Description*	*Example*
-n	optional. specifies how many output lines should be generated	-n=52000
-e	optional. specifies how often the number of generated rows will be displayed.	-e=1000
-c	required. path to the folder where the category files are located	-c=/home/generator/categories
-l	required. name and path of the file, describing the output format	-l=/home/generator/rowlayout.xml
-o	optional. name and path of the file which contains the generated data	-o=/home/generator/output.txt
-v	optional. output status during processing	-v
-f	optional. specifies the output format: 0=regular (mixed) case, 1=lowercase only or 2=uppercase only	-f=2
-p	optional. defines which character set should be used when generating random data	-p=ABCDEFGHIJKLMNOPQRST1234567890!$+*%&
-m	optional. Defines the maximum year used during date generation. Default is 2199	-m=2016
-i	optional. Defines the miximum year used during date generation. Default is 1970	-i=2013

Running the program

Running the program is straightforward. As discussed above, either specify the required arguments or define the arguments in a properties file. As a prerequisite you need to have Java installed on your system.

When you use a properties file then run it as follows:

```
java -cp .:datagenerator.jar com.datamelt.datagenerator.DataCreator
```

When you specify all arguments on the command line:

```
java -cp .:datagenerator.jar com.datamelt.datagenerator.DataCreator
-c=/home/generator/categories -l=/home/generator/rowlayout.xml -v
```

Make sure that you adjust the above parameters according to your needs. If the *datagenerator.jar* file is located somewhere else on your filesystem, add the appropriate path.

Starting with version 0.30, the program first reads the properties file and then parses the command line arguments passed to the program. This allows to define values in a properties file and then to override the relevant parameter by specifying it on the command line when running the program.

Note: The GNU interpreter for Java bytecode - gij – does not work properly with random values. You should use the Oracle (Sun) Java implementation instead.

XML formatting rules

The program does some checks on the xml file that defines the way how the relevant fields and the output data should look like. Anyway, the program requires that you correctly define fields and reference fields. Listed below are some rules that need to be followed:

- Definition: “reference fields” are fields that are not output but are used as a reference for other fields. “regular fields” are fields for which the value will be written to the output file.
- Reference fields need to have a unique “id” attribute specified
- Reference fields may reference other reference fields but make sure no circular references are created.
- Reference fields of type “random” or “regex” required a value for the “length” attribute
- Reference fields of type “category” need a value for the attribute “category”
- Regular fields always need a value for the “length” attribute
- Regular fields need to be of type “category”, “datetime”, “regex”, “random” or “reference”
- Regular fields of type “category” need an attribute “category” defined
- Regular fields of type “datetime” need a pattern specified according to the pattern syntax as defined in the java class SimpleDateFormat. Additionally “q” or “Q” and “h” or “H” may be specified as a pattern value (but not in combination with other pattern characters). “q” will return an integer value for the relevant quarter (e.g. “4”). “Q” will

return the character “q” plus the relevant value for the quarter (e.g. “q3”). “h” will return the value for the relevant half year (e.g. “1”) and “H” will return the character “h” plus the relevant value for the half year (e.g. “h1”).

- Regular fields of type “regex” need a pattern specified which is a regular expression. Note the limitations for regular expressions documented elsewhere in this document.
- Regular fields of type “reference” need an attribute “reference” which contains a value corresponding to the “id” value of the referenced field. Within the reference attribute for a regular field, multiple reference fields may be referenced by dividing the referenced id's by a slash character (“/”).

Sample data

Below find some sample data generated by the program. The rowlayout file, which defines the structure of the output data, is as follows:

```
<xml>
        <references>
                <field type="datetime" id="date1" pattern="yyyy-MM-dd"/>
                <field type="datetime" id="year" reference="date1" pattern="yyyy"/>
                <field type="category" id="airport" category="airport_code"/>
        </references>
        <row type="delimited" seperator=";">
                <field type="category" category="seasons" length="20" />
                <field type="datetime" pattern="q" length="1" />
                <field type="random" length="10" />
                <field type="regex" pattern="www.[a-pA-P]{8,16}.com" length="30" />
                <field type="reference" reference="date1" length="10" />
                <field type="reference" reference="airport/-/year" length="20" />
        </row>
</xml>
```

At the top there are three reference fields defined. The first is a general date with a pattern and an “id” of “date1”. The second field “year” references the “date1” field defined before, so it uses the same date, but a different pattern is applied. The third field defines a field of type category which will be used as a reference later.

Next the fields for the output are defined. The first field uses a file “seasons.category” from which values randomly are picked for output. Next comes a datetime field which will output the quarter of the generated date. The third field generates a random value of length=10 and the fourth field generates a value according to the regular expression pattern specified.

At the end there are two fields defined, that reference fields defined in the “references” section at the top of the file. Field 5 reference the field “date1” explained further above. The last field shows how to reference multiple fields: the referenced id's of the references are divided by a slash (“/”). When a specified id is not found/existing the literal value is used. So the “-” in the “reference” attribute is output as is.

Here is now an example of the generated data according to the rowlayout file described above:

```
Spring;1;SyWds5JVSq;www.MfnCDgkE.com;1996-04-02;BSL-1996
Autumn;1;LU28ghZV71;www.IFFfpHPgnGG.com;2014-08-05;BOS-2014
Autumn;4;UOIlxk4Eh9;www.NEKJMfmFLccFOenG.com;2009-06-07;FRA-2009
Summer;1;KpuQtEct95;www.JedgDBFDPd.com;2015-01-08;BOS-2015
Winter;4;yHLhhFF2yl;www.gjgFhHEFeJfgjFk.com;1998-05-09;ZRH-1998
Winter;4;uwUvP66xXh;www.DbPmpnAoN.com;2009-07-16;BSL-2009
Spring;4;MfvK2Jx4Bp;www.nIGKpMPNbaoBp.com;2015-06-17;DEN-2015
Winter;4;ATO9TbvvNp;www.bNklmlpMBFDc.com;1999-05-18;GVA-1999
Spring;4;iClRTRSoGa;www.PigFkGKIgJifj.com;2011-01-15;DEN-2011
Winter;1;EdQ9acCZBI;www.gboppPNaOlpmAfEI.com;2001-04-02;DEN-2001
Summer;4;AEt5Fw27tO;www.MIMJjkkGdC.com;2011-12-28;GVA-2011
Autumn;4;93eEEpqxkJ;www.MJKKhlilmILLMM.com;2016-07-05;BSL-2016
Summer;4;i48C66hBQS;www.pdAfBFDcEdgeDfE.com;2010-05-11;ZRH-2010
Summer;4;w3duceXBMD;www.odADEaFb.com;2008-09-14;DEN-2008
Spring;4;UjbnVPQt8f;www.PJjkiFkKOOkodgh.com;2011-12-25;KBP-2011
Winter;4;gZN4LExtpS;www.JfKghhGlHIJhEaa.com;2011-08-11;BSL-2011
Spring;4;8zRKaR5cEG;www.aanPompLANKNkOl.com;2018-07-09;GVA-2018
Autumn;1;j8MjWjmWLY;www.ALJiCcPdpdCGD.com;2011-12-24;ZRH-2011
Spring;1;6R4xprupbZ;www.OanlomLME.com;2019-09-06;FRA-2019
Spring;4;FzcUivaOqY;www.aNkimLIM.com;2018-05-14;GVA-2018
Spring;4;RSyMVOgYjq;www.jgGHdnMK.com;2009-01-01;FRA-2009
Summer;4;3YJKTIuPeo;www.domAMBNOOolppAP.com;1996-09-05;BSL-1996
Summer;4;6RSJjRni1i;www.IdeffCFDGEdge.com;2011-12-14;BSL-2011
Spring;4;jyFiPTUNP8;www.cfCpDabb.com;2001-10-02;PFO-2001
Winter;1;XUXAba2BUj;www.JnInKHJihi.com;2006-09-02;GVA-2006
Autumn;4;IjJFQvCngt;www.NjhFhECFbaeJg.com;2019-12-21;GVA-2019
Spring;4;TVf2YkxwV9;www.KHfgOlNkimJHLhi.com;2017-12-06;DEN-2017
Summer;4;9A2oRfEA4z;www.HEGdboaomJNLiDca.com;2000-10-03;GVA-2000
Summer;4;yaTsOFSBK9;www.GcIMKgfiFDH.com;2018-08-31;ZRH-2018
Winter;4;H3zEkyDXqM;www.gBABNLPiiEFo.com;2013-06-22;PFO-2013
```

Alphabetical Index