

DATA GENERATION

CLAIM OF PRIORITY

This application claims priority to U.S. Patent Application Serial No. 61/917,727, filed on December 18, 2013, the entire contents of which are incorporated herein by
5 reference.

BACKGROUND

This description relates to data generation.

During development of data processing applications, developers can work outside of a production environment and may not have access to production data. To ensure that a
10 data processing application will run correctly in production with actual data, realistic data can be used during development and testing of the data processing application.

SUMMARY

In a general aspect, a method includes receiving data indicative of a number of times each of one or more rules was executed by a data processing application during
15 processing of one or more records; based on the number of times each of the rules was executed by the data processing application, determining a content criterion for each of one or more particular fields; generating content for each of the particular fields based on the content criterion; and populating each of the particular fields with the generated
content.

20 Embodiments may include one or more of the following features.

Generating content includes generating content based on a format specification for each of the particular fields. The format specification includes a type of content, including one or more of a string, a decimal, an integer, a date, or a datetime. The method includes receiving the format specification through a user interface.

25 The method includes creating a new record including one or more of the particular fields; and populating the particular fields for the new record.

Populating each of the particular fields with the generated content includes modifying pre-existing content for one or more of the particular fields.

Populating each of the fields includes creating the particular field for one or more pre-existing records.

The content criterion includes one or more of a target value, a target range of values, a target distribution of values, a maximum value, or a minimum value.

5 The content criterion includes particular values for the generated content.

The content criterion is indicative of a target combination of content to be included in two or more of the one or more fields.

The method includes receiving the content criterion through a user interface.

Generating content includes generating content based on a generation mode.

10 The generation mode includes one or more of (i) generation of random content for one or more of the fields for each of multiple records, (ii) generation of unique content for one or more of the fields for each of multiple records, and (iii) selection of content for one or more of the fields for each of multiple records from a set of available content.

The method includes providing one or more records including the generated
15 content to the data processing application.

The content criterion is determined to cause a particular rule to be executed by the data processing application.

Determining the content criterion includes identifying a particular rule that was executed less than a minimum threshold number of times; and determining the content
20 criterion to cause the particular rule to be executed at least the minimum threshold number of times.

The format specification includes a type of content.

In a general aspect, a method includes receiving, through a user interface, a format specification; receiving, through a user interface, a content criterion; receiving,
25 through a user interface, an indication of a mode by which content is to be generated; generating content according to the indicated mode, wherein the generated content satisfies the format specification and content criterion; and creating a field in a record, the field including the generated content.

Embodiments may include one or more of the following features.

The mode includes one or more of (i) generation of random content for each of multiple records, (ii) generation of unique content for each of multiple records, and (iii) selection of content for each of multiple records from a set of available content.

Receiving the content criterion includes receiving a selection of the content
5 criterion from one or more content criteria displayed on the user interface.

Receiving the format specification includes receiving a selection of the format specification from one or more format specifications displayed on the user interface.

The format specification indicates a type of content, such as one or more of a string, a decimal, an integer, a date, or a datetime.

10 The content criterion indicates one or more of a maximum value for the content, a minimum value for the content, and a range of values for the content.

The content criterion indicates particular values for the content.

The content criterion indicates a target distribution for the content.

The content criterion is indicative of a target combination of content to be
15 included in two or more fields.

Receiving the indication of the content generation mode includes receiving a selection of the content generation mode from one or more content generation modes displayed on the user interface.

The method includes creating a new record. Creating a field in a record includes
20 creating a field in the new record.

The method includes receiving a second format specification, a second content criterion, and an indication of a second mode; generating second content according to the indicated second mode, wherein the generated content satisfies the second format specification and the second content criterion; and replacing the content for an existing
25 field in the record with the generated second content.

In a general aspect, a non-transitory computer readable medium stores instructions for causing a computer system to, based on a number of times each of one or more rules was executed by a data processing application during processing of one or more records, determine a content criterion for each of one or more fields; generate
30 content for each of the fields based on the content criterion; and populate each of the fields in one or more records with the generated content.

In a general aspect, a computing system includes at least one processor configured to, based on a number of times each of one or more rules was executed by a data processing application during processing of one or more records, determine a content criterion for each of one or more fields; generate content for each of the fields based on
5 the content criterion; and populate each of the fields in one or more records with the generated content.

In a general aspect, a computing system includes means for, based on a number of times each of one or more rules was executed by a data processing application during processing of one or more records, determining a content criterion for each of one or
10 more fields; means for generating content for each of the fields based on the content criterion; and means for populating each of the fields in one or more records with the generated content.

In a general aspect, a method for testing a data processing application during development of the data processing application includes processing first data records
15 using the data processing application, the data processing application implementing one or more rules. Each first data record has multiple fields. The method includes comparing a number of times each rule of the data processing application is executed during processing of the first data records with a target value for the rule. The method includes
20 based on the results of the comparing, determining a content criterion indicative of a target characteristic for data in second data records to be processed using the data processing application. The method includes generating content for one or more fields of the second data records according to a format of the first data records and according to the content criterion; processing the second data records using the data processing
25 application; comparing a number of times each rule of the data processing application is executed during processing of the second data records with the target value for the rule and determining that the target value for each rule is satisfied; and testing the data processing application using the second data records.

Embodiments can include one or more of the following features.

The method includes receiving the first data records at a client device, wherein the
30 data processing application is executed on the client device. The client device comprises a computing device.

The method includes receiving the first data records from a remote server. The method includes receiving the first data records at a client device that is communicatively coupled to a server via a LAN or WAN connection

5 The method includes determining the content criterion such that the data in the second data records causes the target value for each rule to be satisfied when the data processing application processes the second data records.

Generating content for one or more fields of the second data records includes modifying content in one or more fields of the first data records, and wherein the modified first data records are the second data records.

10 Generating content for one or more fields of the second data records includes creating a new data record.

The method includes storing the second data records in a storage device.

The method includes modifying the data processing application based on a result of the testing.

15 Testing the data processing application includes modifying the data processing application based on a result of processing the second data records using the data processing application.

The method includes determining the content criterion such that the data in the second data records is substantially similar to data expected to be processed by the data processing application in a production environment.

20 Aspects can include one or more of the following advantages. Data can be generated according to arbitrary format specifications and content criteria, e.g., through a straightforward user interface. Data generation can be used to supplement existing data, e.g., if an existing data set does not have enough records or does not have records with
25 desired characteristics. Data generation can also be used to generate data from scratch, e.g., if an application calls for data that is not available.

In some examples, realistic data can be generated for testing of a data processing application. For instance, a set of test data can be generated that causes all of the rules of the data processing application to execute at least once, e.g., that triggers every possible
30 action in the data processing application. If such data do not exist, the data can be generated. For instance, once a complete set of data is generated, the generated set of data

can be processed by the data processing application to test the execution of the data processing application. Because the generated set of data causes every possible action in the data processing application to be executed, the data processing application can be thoroughly tested. If the data processing application does not perform as expected (e.g., if
5 an action of the data processing application does not perform as expected), the data processing application can be modified as appropriate and tested until satisfactory results are achieved.

Other features and advantages of the invention will become apparent from the following description, and from the claims.

DESCRIPTION OF DRAWINGS

FIG. 1 is a block diagram of a system for data generation.

FIG. 2 is a block diagram of a data generation module.

FIG. 3A is a portion of an example source file.

FIGS. 3B-3D are portions of example target files.

FIG. 4 is a source window.

FIG. 5A is a source window.

FIG. 5B is a data generation window.

FIG. 6 is a target window.

FIG. 7 is an auxiliary source window.

FIG. 8 is a record format window.

FIG. 9 is a string characteristics window.

FIG. 10 is a decimal characteristics window.

FIG. 11 is an integer characteristics window.

FIG. 12 is a date characteristics window.

FIG. 13 is a datetime characteristics window.

FIGS. 14A-14C are windows for rule creation.

FIGS. 15A-15C are windows for rule creation.

FIG. 16 is a window for rule creation.

FIGS. 17A-17B are windows for subject-area datasets.

FIG. 18 is a flow chart.

FIG. 19 is a block diagram of a data processing application.

FIG. 20 is a flow chart.

DESCRIPTION

We describe here a general approach to generating data according to various configurable options. For instance, the data type of the generated data can be specified by a user. The data types can include, e.g., string, decimal, integer, date, and datetime. The generated data can satisfy one or more configurable content criteria that impose limits on characteristics of the generated data, such as, e.g., a range of allowed values for generated decimal or integer data, an average string length for generated string data, a set of values or characters that can be used in the generated data, and other characteristics. Data generation can include modifying values in one or more fields of existing source records, augmenting source records by creating and populating new fields in the records, or creating entirely new records. In some examples, the configurable options can be specified by a user through a user interface.

In some examples, a set of records can be processed by a data processing application. Data processing applications can implement rules whose execution depends on the value of one or more variables, such as depending on the input records. In some cases, data can be generated for the set of records to cause a desired execution result when the data processing application processes the set of records. For instance, data can be generated for the set of records that will cause a particular rule in the data processing application to execute, to cause a particular rule to be executed more or fewer times, or to cause another desired execution result. Based on a result of executing the application, a determination can be made as to what data are to be generated, e.g., to better achieve the desired execution result. For instance, a set of realistic data can be generated that is substantially similar to production data that is expected to be processed by the data processing application. The generated set of realistic data can be used to test the data processing application, e.g., to ensure that all aspects of the data processing application operate correctly. Based on the results of the testing, the data processing application can be modified, if appropriate, prior to use in a production environment.

FIG. 1 shows an exemplary data processing system 100 in which the data generation techniques can be used. The system 100 includes a source 102 that may include one or more sources of data such as storage devices or connections to online data

streams, each of which may store or provide data in any of a variety of formats (e.g., database tables, spreadsheet files, flat text files, or a native format used by a mainframe). An execution environment 104 includes an execution module 112. The execution environment 104 may be hosted, for example, on one or more general-purpose computers under the control of a suitable operating system, such as a version of the UNIX operating system. For example, the execution environment 104 can include a multiple-node parallel computing environment including a configuration of computer systems using multiple central processing units (CPUs) or processor cores, either local (e.g., multiprocessor systems such as symmetric multi-processing (SMP) computers), or locally distributed (e.g., multiple processors coupled as clusters or massively parallel processing (MPP) systems), or remote, or remotely distributed (e.g., multiple processors coupled via a local area network (LAN) and/or wide-area network (WAN)), or any combination thereof.

The source 102 includes storage devices that may be local to the execution environment 104, for example, a storage medium connected to a computer hosting the execution environment 104 (e.g., hard drive 108), or may be remote to the execution environment 104, for example, being hosted on a remote system (e.g., mainframe 110) in communication with a computer hosting the execution environment 104, over a remote connection (e.g., provided by a cloud computing infrastructure).

The execution module 112 reads and processes data read from the source 102. The execution module 112 includes one or more data processing applications that execute processes, such as computations, on the data. The output data 114 may be stored back in the source 102 or in a data storage system 116 accessible to the execution environment 104, or otherwise used.

The data storage system 116 is also accessible to a development environment 118 in which a developer 120 is able to configure the data processing applications, sources 102, or both. The development environment 118 is, in some implementations, a system for developing applications as dataflow graphs that include vertices (representing data processing components or datasets) connected by directed links (representing flows of work elements, i.e., data) between the vertices. For example, such an environment is described in more detail in U.S. Publication No. 2007/0011668, titled "Managing

Parameters for Graph-Based Applications,” incorporated herein by reference. A system for executing such graph-based computations is described in U.S. Patent 5,966,072, titled “EXECUTING COMPUTATIONS EXPRESSED AS GRAPHS.” Dataflow graphs made in accordance with the system 100 provide methods for getting information into and out of individual processes represented by graph components, for moving information between the processes, and for defining a running order for the processes. The system 100 includes algorithms that choose interprocess communication methods from any available methods (for example, communication paths according to the links of the graph can use TCP/IP or UNIX domain sockets, or use shared memory to pass data between the processes).

A data generation module 150 can generate data for processing by the execution module. The data generation module 150 can modify or augment existing data, such as data from the source 102. For instance, the data generation module 150 can modify values for one or more fields in source data records or can create and populate one or more new fields in the source data records. The data generation module 150 can also generate entirely new data records, e.g., based on a format of the source data records or based on a specified format.

Referring to FIG. 2, the data generation module 150 includes a data generation engine 20 that generates data to be stored in a target 24. In some examples, the data generation engine 20 can generate data based on the data in a data source 22. In some examples, the data generation engine 20 can generate data from scratch, i.e., without source data. The data source 22 can be a file, a database, or a parameter set, or another source of data. The data source 22 can include one or more records, each having one or more fields of data. For instance, an example data source can be a database storing customer records for customers of a retail store (which we refer to here as the “customer database”). Each record in the database represents an individual customer. Each record can have multiple fields. The data source 22 can have a record format that specifies the format of the records, such as the number of fields, the type of data in each field (e.g., string, integer, decimal, date, or datetime), and characteristics of the data in each field (e.g., an allowed range of values, a maximum allowed value, a list of allowed characters,

or other limitations). In some examples, data is generated from scratch and no data source is provided.

For instance, referring to FIG. 3A, a portion of an example source file 200 includes six records 202, each record having the following fields: a *cust_ID* field 204, a *name* field 206, a *cust_age* field 208, and a *state* field 210.

Referring again to FIG. 2, in some examples, the data source 22 can be identified by a user 26. For instance, a user interface engine 28 of the data generation module 150 can enable a user interface 30 to be displayed to the user 26, e.g., on a display interface of a computing device 32, such as a personal computer or a mobile computing device. The user 26 can identify the data source 22 through the user interface 30. The generated data can be stored in the target 24, such as a file, a database, a parameter set, or another data structure. In some examples, the target file 24 can be identified by the user 26, e.g., through the user interface 30. In some cases, the record format of the target can match the record format of the source. That is, a record format that is specified for the target can be the same as the record format of the existing records in the source. In some cases, the record format of the target differs from the record format of the source, and some or all of the record formats of the source can be mapped to the target. For instance, the format of one or more fields of the source can be included in the record format of the target.

The data generation module 150 can generate data for the target 24 based on configuration data 34, e.g., stored in a database 36, a file, or another data structure. The configuration data 34 can specify a data generation approach to be used for the target 24, a content generation mode for the target 24, a data type of the data to be generated for the target 24, content criteria for the data to be generated for the target, and other configuration information for the data to be generated for the target 24. These configuration data 34 are discussed in greater detail below. In some cases, some or all of the configuration data 34 used to generate the data for the target 24 can be specified by the user 26, e.g., through the user interface 30. In some examples, some or all of the configuration data 34 can be determined by a configuration engine 39, for instance, based on an analysis of the data source 22 or based on information about desired properties of the target.

In some examples, the data generation engine 20 can generate data for the target 24 by modifying values for one or more of the fields of existing source records in the data source 22 according to the configuration data 34. The modified records can be stored in the target 24. In some cases, all of the values for a given field can be modified. For instance, a value can be assigned to a given field for each record such that the distribution of values in the given field across all of the records matches a target distribution as indicated by the configuration data 34. For instance, the distribution can be specified by the user 26 or determined by the configuration engine 39. In some cases, fewer than all of the values for a given field can be modified, such as only values that do not meet a specified criterion as indicated by the configuration data 34. For instance, any values for a given field that fall outside of a particular range of allowed values for that field can be modified.

For instance, referring to FIG. 3B, a portion of an example target file 220 includes records 222 generated based on the source file 200. In this example, the configuration data 34 (FIG. 2) indicate that the *cust_age* field 208 can have a maximum value of 40. To satisfy the criterion specified by the configuration data 34, the values for the *cust_age* field 208 in the source file 200 were modified by the data generation engine 20 such that the maximum value for the *cust_age* field was 40. That is, the value for the *cust_age* field 208 for the particular record 222a was modified from 45 (in the source file 200) to 25 (in the target file 220). Values for the other fields of the generated records 222 of the target file correspond to the values in the source records 202.

Referring again to FIG. 2, in some examples, the data generation engine 20 can generate data by augmenting the existing source records of the data source 22 with one or more new fields according to the configuration data 34. The augmented records (i.e., the records with the one or more new fields) can be stored in the target 24. The configuration data 34 can provide instructions for determining the number of new fields, the data types and values for the new fields, and other characteristics of the new fields. In some cases, the configuration data 34 can specify that values for a new field are to be generated based on the data for an existing field in the source data of the data source 22. In some cases, the configuration data 34 can specify that values for a new field are to be generated from

scratch, e.g., according to specified characteristics (e.g., characteristics specified by the configuration data 34) but not based on any existing source data.

For instance, referring to FIG. 3C, a portion of an example target file 230 includes records 232 generated based on the source file 200. In this example, the configuration data 34 (FIG. 2) specify the creation of a new *age_range* field 234 that has values that are generated by classifying the value for the *cust_age* field 208 into one of five available *age_range* buckets. The configuration data 34 further specify the creation of a new *cust_income* field 236 with values subject to specified maximum and minimum values but without any relationship to the data in the source file 200.

Referring again to FIG. 2, in some examples, the data generation engine can generate data by augmenting the existing source records of the data source 22 with one or more new records according to the configuration data 34. The augmented records (i.e., both the existing source records and the new records) can be stored in the target 24. The new records can have the same record format as the source records. The configuration data 34 can provide instructions for determining the number of new records, the values for the fields of the new records, and other characteristics of the new records. In some cases, the configuration data 34 can specify that values for one or more fields in the new records are to be generated from scratch, e.g., according to specified characteristics, but not based on any existing source data of the data source 22. In some cases, the configuration data 34 can specify that values for one or more fields in the new records are to be generated to satisfy a profile specified by the configuration data 34. For instance, the profile can specify that the values for a particular field in all of the records satisfy a characteristic, such as an average or a distribution. For instance, in the customer database source, new records can be generated such that the values for the *cust_age* field across all of the records satisfy a particular distribution specified by the configuration data 34.

For instance, referring to FIG. 3D, a portion of an example target file 240 includes records 202, 242 generated based on the source file 200. In this example, the configuration data 34 (FIG. 2) specify the creation of four new records 242 such that the values for the *cust_age* field 208 across all of the records have an average value between 40 and 45, subject to a maximum value of 60 for the field. In this example, the configuration data 34 specify that the values for each other field are to be randomly

generated, e.g., according to specified characteristics for the field. The original source records 202 are also included in the target file 240.

In some examples, the data generation engine 20 can apply more than one approach to data generation, according to the configuration data 34. For instance, the data generation engine 20 can apply any combination of the following approaches: the values for one or more fields can be modified, the source records can be augmented with one or more new fields, and the source records can be augmented with one or more new records.

In some examples, only generated records are stored in the target, such as a file, a database, or a parameter set, and the source records (if any) are not stored in the target. In some cases, a source can be specified, e.g., by a user, and the generated records can be generated by the data generation engine 20 module based on the record format of the source, a profile of one or more fields of the source, or another characteristic of the source. In some cases, no source is specified and the records are generated automatically from scratch by the data generation engine 20 according to the configuration data 34. For instance, the record format and the characteristics of one or more fields can be specified by the configuration data 34.

In some examples, the record format of the source can be mapped to the target. For instance, the configuration data 34 can indicate that the record format of the source is to be adopted by the target. For instance, the record format of the source can be applied to the target according to the configuration data 34 and new records can be generated from scratch by the data generation engine 20 according to the record format of the source. In some cases, data can be generated from multiple sources, and the record format of each source can be partially or completely mapped to the target. For instance, the format of fields of interest from each source can be mapped to the target.

In some examples, the record format of the source can be mapped to the target and modified. For instance, the configuration data 34 can specify that the name of a field is to be changed from the source to the target, or that a field from the source is to be removed.

Referring to FIG. 4, the data generation module 150 provides a user interface (e.g., the user interface 30 of FIG. 2) that provides a source window 300 to enable a user to identify the data source. The source window 300 includes a source type menu 302 that

allows the user to specify a source type, such as a file or a database. In the example of FIG. 4, the source is a file. The source window 300 enables the user to specify an identifier 304 of the source, such as a path to the source (e.g., a path to a file source or a path to a configuration file for a database source) and an identifier of a record format of the source records (e.g., a path to a record format file). In some examples, when the source is a database, the user can specify a query (e.g., a SQL query) that is to be used to obtain source data from the database. The source window 300 provides an option to allow the user to indicate whether new records 306 are to be generated, and if so, the number of new records to be generated. The source window 300 enables the user to view other information 308 about the source. For instance, the user can view the record format of the source, view the source data, or view a profile of the source data. In some examples, an executable application, such as a dataflow graph, can be used to generate source data.

Referring to FIGS. 5A and 5B, in some examples, the source window 300 of the user interface 30 allows the user to specify that data is to be generated from scratch without specifying a source. In particular, the source window 300 enables the user to select generated data as the source type in the source type menu 302. If the user selects generated data as the source type, a data generation window 400 is displayed in the user interface 30. The data generation window 400 enables the user to indicate a method 402 to be used to generate the data (discussed below) and indicate a number of new records 404 to be generated. In some examples, default data generation characteristics, such as a default data generation mode or default values that are specified in the record format of the target can be used to generate data from scratch unless the user overrides the default. For instance, a field generation overrides option 406 allows the user to specify data generation characteristics for one or more fields in the target. In some examples, a rules option 408 allows the user to create more complex data generation rules for individual fields.

Referring to FIG. 6, the user interface 30 provides a target window 500 that enables the user identify the target. A target type menu 502 in the target window 500 enables the user to specify the type of the target, such as a file or database. The target window 500 also enables the user to specify an identifier 504 of the target (e.g., a path to a target file or a path to a configuration file for a target database). In some examples, the

record format of the source is used as the record format for the target unless a different record format is specified, e.g., through the target window 500. For instance, if data is generated from scratch without a source, a record format file can be identified through the target window that specifies the record format for the target. In some examples, an executable application, such as a dataflow graph, can be used to write the generated data to the target. The target window 500 provides a run button 506 that provides the user with access to various configurable options for data generation once the source and target have been identified. The target window 500 also enables the user to view other information 508 about the source. For instance, the user can view the target data set, view a profile of the target data set, or view the record format of the target. In some examples, the target window provides access to statistics associated with the generation of data for the target, such as the number of records read from the source and/or written to the target, the number of records generated, or other statistics.

The data generation engine 20 provides several approaches to generate data. A data generation window of the user interface 30 enables the user to specify the desired data generation approach. For instance, field modification, field creation, record creation, existing source, and parent dataset approaches can be used to generate data.

Field modification: In a field modification approach, the values for one or more fields of the source records can be modified, such that the target data records have the same format as the source data records but different content. In some cases, all of the values for a given field can be modified. For instance, values can be assigned such that the distribution of values in a given field across all of the records matches a target distribution. In some cases, fewer than all of the values for a given field can be modified, such as only values that do not meet a specified criterion. For instance, any values that fall outside of a particular range of allowed values for a particular field can be modified.

Field creation: In a field creation approach, one or more new fields can be created for existing records. In some cases, values for a new field can be generated based on the data for an existing field in the source data. In some cases, values for a new field can be generated from scratch, e.g., according to specified characteristics but not based on any existing source data.

Record creation: In a record creation approach, new records can be generated. When new records are to be generated, the number of records to be generated can be specified (e.g., as in FIG. 5B). In some cases, the record format of the new records can be specified. For instance, if the target is to be populated with both existing source records and newly generated records, the record format of the new records can be the same as the record format of the source records. If the target is to be populated with only newly generated records, the user can specify the record format to be applied to the generated records, such as the number of fields, the type of data for each field, the characteristics of the data for each field (e.g., a maximum value, a minimum value, a set of allowed characters, and other characteristics), and other features of the record format.

Existing dataset: In an existing dataset approach, data is generated for a newly created child dataset. For instance, based on an existing customer dataset, a new dataset including transaction records can be generated. In the existing dataset approach, a specified number of new records are generated for each existing source record. For instance, referring to FIG. 7, an existing dataset window 600 allows a user to specify a path 602 or data manipulation language (DML) 604 of the source dataset, a number 606 of target records to be generated per data record in the existing source dataset, and a percentage of source records that are not to be used to generate records in the target. In some examples, an auxiliary source can contain data to be used to populate certain fields of target records but can have a record format that does not match the record format of either the source or the target. In such cases, one or more fields of interest from the auxiliary source can be mapped to the target records.

Subject-area approach: In a subject area approach, an existing source dataset can be a parent dataset in a subject-area hierarchy, and generating data can include generating a child dataset that is related to the parent dataset. For instance, the parent dataset (i.e., the source) can be a set of customer records and the child dataset (i.e., the target) can be a set of one or more transaction records for each customer. A key field can be specified that links records in the child dataset to corresponding records in the parent set, and one or more key fields of the parent data records are mapped to corresponding fields in the generated child data records. For instance, the *cust_ID* field can be the key field linking customer records and transaction records. The number of child records to be generated

per key field can be specified. In some cases, the percentage of parent records that are not to be used to generate child records can be specified. In some cases, the record format for the child records can be specified.

In some examples, data can be generated according to a format specification, a content criterion, or both. A format specification is a specification of the format of the data to be generated. For instance, the format specification can indicate the data type (e.g., string, decimal, integer, date, or datetime) of the data to be generated. The content criterion is a criterion that limits the characteristics of the data to be generated. Example content criteria can include, e.g., an allowed range of values, a maximum allowed value, a list of allowed characters, or other content criteria. In some examples, the format specification and the content criterion are specified in the record format of the target records. In some examples, the user interface 30 can provide field windows that enable the user to specify characteristics of a field, such as a format specification or a content criterion for the field.

Referring to FIG. 8, a record format window 700 of the user interface enables a user to edit the target record format, including the data type and the data characteristics for one or more fields of the target. The record format window 700 displays a list 702 of the fields that are in the record format of the target. The list 702 also indicates the data type for each field. In some examples, fields that are in the record format of the target also appear in the record format of the data source. Those fields that appear in both the target record format and the source record format can be marked in the list 702, e.g., with asterisks (i.e., **). For instance, in the example of FIG. 8, the fields *last_transaction* and *customer_since* appear in both the target record format and the source record format. The fields that are not marked appear only in the target record format. In some examples, fields that appear in the source record format but not in the target record format are not displayed in the list 702.

The record format window 700 enables a user to select one or more fields of the target record format for which to specify data generation characteristics, e.g., by clicking on, tapping on, or dragging the name of the desired fields. A list 704 of the selected fields of the target record format is displayed. That is, the fields in the list 704 are those fields of the target record format for which the user intends to specify data generation

characteristics. In some examples, the list 704 can be a subset of the list 702 of all of the fields in the target record format, e.g., if the user intends to specify data generation characteristics for only some of the fields of the target record format. For instance, in the example of FIG. 8, the fields *custid*, *fname*, *lname*, and *building_num* have been selected.

5 The user interface 30 enables a user to edit the record format for each of the selected fields displayed in the list 704. For instance, for each of the selected fields, the user can perform any combination of designating the data type for the field, assigning a content generation mode to the field, and specifying data characteristics for the field. The user interface can display one or more of a data type window, a content generation
10 window, and a data characteristics window for each of the selected fields in turn, thus enabling the user to specify various features for each of the selected fields.

 The user interface 30 enables the user to designate a data type for the field. A data type window can be displayed in the user interface 30 responsive to user selection of a particular field from the list 704 to enable the user to designate a data type for that field.

15 The user can designate any arbitrary data type that is supported by the data generation module 150, such as strings, decimals, integers, dates, and datetimes (i.e., a date and a time). One data type can be designated for each field. In some cases, if a data type is not designated for a given field, the data type indicated by the record format for that field is retained. For instance, in the example of FIG. 8, the data type of each of the fields as
20 indicated by the record format for the field is shown in parenthesis next to the field name. The data type designated for each field of the target record format can be stored as configuration data 34.

 The user interface 30 enables the user to assign a content generation mode to one or more of the selected fields. A content generation window can be displayed in the user
25 interface 30 responsive to user selection of a particular field from the list 704 to enable the user to assign a content generation mode to that field. The content generation mode for a field indicates the way in which data is to be generated for the field by the data generation engine 20. For instance, example content generation modes can include unique mode, random mode, default mode, and list mode. In unique mode, a unique value
30 can be generated for the selected field in each record. For instance, an ascending set of unique decimal, string, or integer values can be generated. Date values can be generated

ascending by day from a default or specified date basis, such as 19000101 or another date basis. Datetime values can be generated ascending by minute from a default or specified datetime basis, such as the current date and time. In random mode, a random value can be generated for the selected field in each record. In default mode, default characteristics or default values that are specified in the record format of the target can be used to generate values. For instance, in some examples, the default data generation mode can be random mode, and default value ranges and characteristics can apply to each data type. In list mode, values are selected from a list of allowed values. In some cases, if a content generation mode is not assigned for a given field, the default mode can be used to generate data for that field or a default value (e.g., as specified in the record format of the target records) can be used as the value for the field. In some examples, the content generation mode for one or more fields can be specified in the record format of the target records. The content generation mode for each field of the target record format can be stored as configuration data 34.

If unique mode, random mode, or default mode data generation is assigned to a field, the user interface 30 enables the user to specify data characteristics for one or more of the selected fields. The data characteristics can depend on the data type for the field and can limit the allowable values that can be generated for the field. For instance, data characteristics for a numerical field, such as a decimal field or an integer field can indicate a maximum allowed value, a minimum allowed value, an allowed range of values, or an average value for the field across all records. Other data characteristics can also be applied, as discussed below. In some examples, a field error rate (i.e., a number or percentage of records to contain errors) can also be specified for one or more fields. To enable data characteristics to be specified for a particular field, a characteristics window that is specific to the data type of that field is displayed in the user interface 30 responsive to user selection of that field from the list 704. For instance, a string characteristics window, a decimal characteristics window, an integer characteristics window, a date characteristics window, or a datetime characteristics window, or a window for another data type can be displayed in the user interface 30. The data characteristics for each field of the target record format can be stored as configuration data 34.

Referring to FIG. 9, a string characteristics window 800 allows the user to specify data characteristics for a string field. In the example of FIG. 9, the data characteristics include an average string length 802, a maximum string length 804, and a set of allowable characters 806 for the particular string field. The user can also specify a percentage of NULL values 808 for the particular string field. In some cases, other data characteristics can also be specified. In some examples, if the user does not specify one or more of the data characteristics for a string field, a default value for the data characteristic can apply. For instance, the default average string length 802 can be 8; the default maximum string length 804 can be 16; the default set of allowable characters 806 can include the characters {a-z, A-Z, 0-9, and space}; and the default percentage of NULL 808 values can be zero.

In one particular example, a user specified the average string length 802 as 4 characters, the maximum string length 804 as 8 characters, the set of allowed characters 806 as {a-z, A-Z}, and the percentage of NULL values 808 as zero for a particular field. In this example, if the user assigned random mode data generation to the particular field, a random string would be generated for that field for each record from the set of allowable characters, subject to the specified average length and maximum length. If the user assigned unique mode data generation to the particular field, a unique string would be generated for that field for each record (i.e., such that no two records would have the same generated string for the particular field).

Referring to FIG. 10, a decimal characteristics window 900 allows the user to specify data characteristics for a decimal field. In the example of FIG. 10, the data characteristics include a range of allowed values 902 and a percentage of NULL values 904 for the particular decimal field. In some cases, other data characteristics can also be specified, such as a maximum number of allowed decimal places, an average value for the decimal field of all the generated records, or other characteristics. In some examples, if the user does not specify one or more of the data characteristics, a default value can apply. For instance, the default range 902 can be 0 to 1000 and the default percentage of NULL values 904 can be 0.

In one particular example, a user specified the range of allowed values 902 as 0-10 and the percentage of NULL values 904 as 10% for a particular field. In this example,

if the user assigned random mode data generation to the particular field, a random decimal that falls within the range of allowed values 902 would be generated for that field for each record. Ten percent of the records would be given a NULL value. If the user assigned unique mode data generation to the particular field, a unique decimal that
5 falls within the range of allowed values 902 would be generated for that field for each record (i.e., such that no two records would have the same value for the particular field).

Referring to FIG. 11, an integer characteristics window 180 allows the user to specify data characteristics for an integer field. In the example of FIG. 11, the data characteristics include a range of allowed values 182. In some cases, other data
10 characteristics can also be specified, such as a percentage of NULL values 184 for the particular integer field or an average value for the integer field of all the generated records. In some examples, if the user does not specify one or more of the data characteristics, a default value can apply. For instance, the default range 182 can be 0 to 127.

Referring to FIG. 12, a date characteristics window 190 allows the user to specify data characteristics for a date field. In the example of FIG. 12, the data characteristics include a date range metric 192, such as years, months, or days. The data characteristics also include a date basis 194, which is the date to use as the basis relative to which a date range is to be calculated (e.g., written as YYYYMMDD), where YYYY refers to the
20 year, MM refers to the month, and DD refers to the day. The data characteristics also include a range of allowed dates 196 relative to the date basis 194 and a percentage of NULL or blank values 198 for the particular date field. In some cases, other data characteristics can also be specified. In some examples, if the user does not specify one or more of the data characteristics, a default value can apply. For instance, the default
25 date range metric 192 can be years, the default date basis 194 can be today, the range of allowed dates 196 can be -7 to 0 years (relative to the default date basis), and the default percentage of NULL values 198 can be 1%.

In one particular example, a user specified the date range metric 192 as days, the date basis 194 as January 1, 2013, the range of allowed dates 196 as 365 (i.e., 365 days
30 beyond the date basis of January 1, 2013), and the percentage of NULL values 198 as zero for a particular field. In this example, if the user assigned random mode data

generation to the particular field, a randomly selected date between January 1, 2013, and January 1, 2014, would be assigned to that field for each record. If the user assigned unique mode data generation to the particular field, a unique date between January 1, 2013, and January 1, 2014, would be assigned to that field for each record (i.e., such that
5 no two records would have the same date for the particular field).

Referring to FIG. 13, a datetime characteristics window 250 allows the user to specify data characteristics for a datetime field. In the example of FIG. 13, the data characteristics include a datetime range metric 252, such as years, months, or days. The data characteristics also include a datetime basis 254, which is the date and time to use as
10 the basis relative to which a datetime range is to be calculated (e.g., written as YYYYMMDD HH24MISS.NNN), where YYYY refers to the year, MM refers to the month, DD refers to the day, HH24 refers to the hour on a 24-hour clock, MI refers to the minute, SS refers to the second, and NNN refers to the thousandth of a seconds. The data characteristics also include a range of allowed datetimes 256 relative to the datetime basis
15 254 and a percentage of NULL values 258 for the particular datetime field. In some cases, other data characteristics can also be specified. In some examples, if the user does not specify one or more of the data characteristics, a default value can apply. For instance, the default datetime range metric 252 can be days, the default datetime basis 254 can be now, the default range of allowed datetimes 256 can be -7 to 0 days, and the
20 default percentage of NULL values 258 can be 0.

In some examples, one or more of the characteristics windows (e.g., the string characteristics window 800, the decimal characteristics window 900, the integer characteristics window 180, the date characteristics window 190, or the datetime characteristics window 250) enables a user to override a default error rate for a field, e.g.,
25 by selecting or specifying a value for a field error rate for the field. For instance, the user can select from the following error rates: none, 1 in 2 fields in error, 1 in 10, 1 in 100, 1 in 1000, 1 in 10000, and 1 in 100000. In some examples, the user can select or specify a percentage error rate.

In some examples, the user interface 30 provides a mechanism for a user to
30 override the default data characteristics for one or more data types. For instance, the user can override the average string length for all string data. When the user overrides a

default characteristic for a particular data type, that override characteristic applies to all data generation of that data type for which the user does not specify a data characteristic.

If list mode data generation is assigned to a field, a list window can be displayed in the user interface 30 that enables a list of allowed values for the field to be provided.

- 5 For instance, if list mode data generation is assigned to the *building_num* field, the user can provide a list of building numbers (e.g., {10, 20, 30, 40, 50}) that can be used as values for the *building_num* field. When data is generated for the *building_num* field by the data generation engine 150, the value of the field for each record can be selected at random from the list of building numbers. In some examples, other constraints can apply.
- 10 For instance, the value of a field can be selected from the list subject to a constraint that each item on the list be selected an equal number of times.

- In some examples, advanced rules for data generation or source-to-target format mapping can be created. For instance, advanced rules can include rules to map values in one field into another field, rules to use the values in one field to calculate the values in another field, or rules to use more than one field in the source to create a single field in the target. In some examples, the advanced rules can relate values in a field of the data source to values to be generated for a field of the target. In some examples, the advanced rule can relate values in a first field of the target to values to be generated for a second field of the target. The user interface 30 can provide a rules editor capability that enables
- 15 a user to generate an advanced rule for generation of data for one or more target fields, such as the configuration screen and rules editor shown in FIGS. 14A-14C, 15A-15C, and 16. For instance, the user interface 30 can provide a rules window that enables the user to specify a rule that is associated with a field, where the rule can be used to generate data to populate the field. The user interface 30 can provide a rules editor that enables the user to
- 20 define the rule.
- 25

- For instance, referring to FIGS. 14A-14C, the rules editor capability of the user interface 30 enables a user to create an advanced rule to map values in one field into another field. In this example, the user wants to create a rule that enables the generation of both a *state_cd* field whose values are selected from a list of two-letter abbreviations
- 30 of the New England states based on a *state_name* field whose values are the corresponding full state names. To create this rule, the user can assign list generation

mode to the *state_cd* field through a configuration window 350 for the *state_cd* field (FIG. 14A) and specify a list 352 (in this case, the list of two-letter abbreviations of the New England states: MA, CT, RI, VT, ME, NH) from which the values for the *state_cd* field can be selected. In a rules window 354 (FIG. 14B), the user can indicate that the *state_name* field 356 is to be associated with a rule “Compute *state_name*” 358 that can be used to populate data in the *state_name* field 356. The user can define the “Compute *state_name*” rule for the *state_name* field in a rules editor 360 (FIG. 14C). In this example, the “Compute *state_name*” rule specifies the value that is to be assigned for the *state_name* field for each value of the *state_cd* field.

For instance, referring to FIGS. 15A-15C, the rules editor capability of the user interface 30 enables a user to create an advanced rule to use the value for a first field in the target to calculate the value for a second field in the target. In this example, the user wants to create a rule that uses the value for an *income* field to calculate the value of a *tax_rate* field. To create this rule, the user can assign random generation mode to the *income* field through a content generation window 450 for the *income* field (FIG. 15A) and specify data characteristics 452, such as a range of allowed values and a percentage of blank values, for the *income* field. In a rules window 454 (FIG. 15B), the user can indicate that the *tax_rate* field 456 is to be associated with a rule “Compute *tax_rate*” 458 that can be used to populate data in the *tax_rate* field 456. The user can define the rule for the *tax_rate* field in a rules editor 460 (FIG. 15C). In this example, the “Compute *tax_rate*” rule 458 specifies the value that is to be assigned for the *tax_rate* field for each value of the *income* field. The “Compute *tax_rate*” rule also specifies a default value 462 to be assigned for the *tax_rate* field, e.g., if the *income* field is blank, NULL, or in error.

For instance, referring to FIG. 16, the rules editor capability of the user interface 30 enables a user to create an advanced rule to map multiple source fields into a single target field. In some cases, the record format of one or more of the source fields can be different from the record format of the target field. In this example, the user wants to create a rule that specifies an expression to use the value of both a *firstname* field and a *lastname* field in generating data for a *fullname* field. To create this rule, in a rules window 550, the user can define a rule 552 for the *fullname* field 554 as a combination of

the *firstname* field 556 from the source, a space, and the *lastname* field 558 from the source (“firstname” + “ ” + “lastname”).

Data can be generated based on one or more of the format specification(s), content criterion or criteria, data generation mode, and advanced rules. For instance, data
5 in the source records can be modified, one or more new fields can be created and populated with data, and/or new records can be created and populated with data.

In some examples, a subject-area can be created. A subject-area is a collection of datasets that are related to each other in a hierarchy through a key relationship in at least one field. For instance, an existing parent dataset can be a set of customer records and a
10 child dataset can be a set of one or more generated transaction records for each existing customer.

Referring to FIG. 17A, in one example, a join window 70 allows a user to specify the one or more fields of the existing parent dataset to be used as a key field with a child dataset to be generated. In the example of FIG. 17A, the existing parent dataset is a
15 customer dataset and a transaction dataset is to be generated as a child with a join key field *custid*. Referring to FIG. 17B, a join window 75 allows a user to specify one or more join key fields for the child dataset. In this example, the *custid* field has already been specified as a join key field from the parent dataset at box 76. At box 78, a field can be specified as a join key field from the child dataset. In some examples, non-joining
20 records, such as a parent record without a corresponding child record or a child record without a corresponding parent record, can be excluded from the subject-area dataset. In some examples, even non-joining records can be included in the subject-area dataset.

In some examples, a subject-area can be created for referential integrity, meaning that the subject-area will provide only data records that ensure referential integrity of the
25 hierarchical datasets of the subject-area. By referential integrity, we mean that that each record of the subject-area is joined to at least one other record of the subject-area. For instance, an example subject-area may include a first dataset that associates an account identifier with each transaction record. In another dataset of the subject-area, each customer identifier may be associated with multiple account identifiers and with multiple
30 product identifiers. Ensuring referential integrity of the subject-area means ensuring that all of the records included in the subject-area datasets are validly joined with one or more

other records in the subject-area dataset. In some examples, a subject-area can be created to include both data records that result from any applicable subsetting rules and data records that ensure referential integrity. That is, for instance, even records that do not satisfy a referential integrity criterion may be included in a subject-area to satisfy a subsetting rule. Further description of subsetting rules is provided in U.S. Patent Application No. 13/827,558, entitled "Data Records Selection," the contents of which are incorporated herein by reference.

In some examples, a virtual dataset can be created based on one or more existing datasets. For instance, based on a customer dataset and a transaction dataset, a new rollup dataset can be generated that includes one or more fields having values such as a number of transactions per customer or a dollar amount spent per customer. Such a rollup dataset can be useful, e.g., in a data processing application that operates on high volume or low volume customers.

Referring again to FIG. 2, in some examples, an analysis engine 38 can perform automated analysis to implement one or more of the format specification(s), content criterion or criteria, data generation mode, and advanced rules for the generated data. For instance, as a basic example, if a user specifies a date basis of May 1, 2000 and a range of allowed dates as 10 years, the analysis engine 38 can identify May 1, 2000, to May 1, 2010, as the range of allowed dates without the user specifying the exact dates.

In another example, the analysis engine 38 can determine the distribution of string lengths across multiple records that will achieve the average string length and maximum string length specified by a user. For instance, the analysis engine can determine the distribution of string lengths based on a default target ratio between a default maximum string length and a default average string length. The default target ratio can be applied to the maximum string length to identify the peak string length in a string length distribution, such as a Gaussian distribution. For instance, in one example, the default maximum string length is 16 and the default average string length is 8, such that the default target ratio is 0.5. This default target ratio can be applied to the generation of strings, e.g., the random generation of strings. For instance, in a field that is limited to a maximum string length of 6, the default target ratio is applied to determine a peak string length of 3. Strings will thus be generated in a Gaussian distribution around the peak

string length of 3. In another example, in a field that is limited to a maximum string length of 25, the default target ratio is applied to determine a peak string length of 12-13. Strings will thus be generated in a Gaussian distribution around the peak string length of 12-13.

5 In some examples, the analysis engine 38 can determine one or more of the format specification(s), content criterion or criteria, data generation mode, and advanced rules automatically, e.g., by an automated analysis of source records. For instance, the analysis engine 38 can determine how to modify existing data and/or generate new data for a particular field to achieve a target distribution of values for that field.

10 Referring to FIG. 18, in a general approach to data generation 650, a format specification is received for a field (652). The format specification can be, e.g., a data type for data to be generated for the field, such as a string, decimal, integer, date, or datetime. For example, the system 100 can generate data according to the general approach to data generation 650. In some examples, the format specification can be
15 received from a user through a user interface, such as user interface 30 (e.g., a data type window provided through the user interface 30).

A content criterion is received for the field (654). The content criterion can be, e.g., a data characteristic for the data to be generated for the field. For instance, data characteristics for a numerical field, such as a decimal field or an integer field, can
20 indicate a maximum allowed value, a minimum allowed value, an allowed range of values, or an average value for the field across all records. Data characteristics for a string field can indicate a maximum allowed number of characters, a minimum allowed number of characters, an average number of characters, or a set of allowed characters. Data characteristics for a temporal field, such as a date field or a datetime field, can
25 indicate a date or datetime metric, a date or datetime basis, or a date or datetime range. In some examples, a field error rate (i.e., a number or percentage of records to contain errors) can also be specified for one or more fields. In some examples, the content criterion can be received from a user through a user interface, such as the user interface 30 (e.g., the string characteristics window 800, the decimal characteristics window 900,
30 the integer characteristics window 180, the date characteristics window 190, or the datetime characteristics window 250).

An indication of a content generation mode for the field is received (656). The content generation mode is the mode by which content is to be generated for the field. For instance, the content generation mode can be selected from the following modes: unique mode, random mode, default mode, and list mode. In unique mode, a unique value
5 can be generated for the selected field. In random mode, a random value can be generated for the selected field. In default mode, default characteristics or default values, e.g., that are specified in the record format of the target, can be used to generate values. In list mode, values are selected from a list of allowed values. In some cases, if a content generation mode is not assigned for a given field, the default mode can be used to
10 generate data for that field. In some examples, the indication of the content generation mode is received from a user through a user interface, such as the user interface 30 (e.g., a content generation window provided through the user interface).

Content for the field for each of multiple records is generated (658) according to the indicated content generation mode. The generated content for the field satisfies the
15 format specification and the content criterion for the field. In some examples, a new field can be created for each of multiple existing records, and content can be generated for the new field according to the indicated content generation mode. In some examples, multiple new records can be created, each record having one or more fields for which content can be generated according to the indicated content generation mode. For
20 instance, multiple new records can be created based on each existing record.

In some examples, data generation can be used to create a set of records that has one or more desired features. For instance, data generation can be used to create a set of records having a desired number of records and/or a desired distribution of values in a field, for processing by a data processing application. Example data processing
25 applications include, e.g., data processing applications for customer billing (e.g., for telephone customer billing), data processing applications that process customer records (e.g., to identify demographic trends in customer purchasing patterns), and other types of data processing applications.

Data processing applications often implement rules whose execution depends on
30 (e.g., is triggered by) the value of one or more variables. For example, the variables may be input variables corresponding to input data, or derived variables that depend on one or

more input variables. For effective testing of the data processing application, input data can be provided that is sufficient to cause every rule in the application to be executed (e.g., such that complete code coverage in the application is achieved), such that every rule is executed at least a corresponding minimum number of times, and/or such that
5 every rule is executed no more than a corresponding maximum number of times.

Referring to FIG. 19, in some examples, input data records 50 can be provided as input data to a data processing application 52. The data processing application 52 implements one or more rules 54. A rule is a set of criteria that can be used, for example, for converting data from one format to another, making determinations about data, or
10 generating new data based on a set of input data. Further description of rules can be found in U.S. Patent Application Serial No. 11/733,434, filed April 10, 2007, the contents of which are incorporated herein by reference in their entirety.

Each rule 54 implemented by the data processing application 52 can be executed when a corresponding condition expression for the rule is satisfied and is not executed by
15 the data processing application 52 if the corresponding condition expression is not satisfied. A rule 54 can be specified by a specification that includes at least a condition expression and an execution expression. When the condition expression is satisfied (e.g., the condition expression is evaluated with a true result), the execution expression can be evaluated. The condition expression can depend on (e.g., is triggered by) the value of one
20 or more variables, which can be, e.g., input variables corresponding to input data 50 or derived variables that depend on one or more input variables. In some cases, the application executes all of the rules 54 that are triggered. In some cases, the application executes fewer than all of the rules 54 that are triggered, such as some of the rules 54 or only one of the rules 54 (e.g., a first rule 54 that is triggered). Further description of the
25 execution of such data processing applications is provided in U.S. Patent Application No. 13/827,558, entitled "Data Records Selection," the contents of which are incorporated herein by reference.

The rules can be any rule that can be executed when a corresponding condition expression is satisfied. In one example, a rule in a data processing application that
30 identifies marketing offers for customers of a retail store can be a rule that identifies a particular marketing offer for customers that have an income less than \$20,000. In one

example, a rule in a data processing application that generates bills for customers of a telephone company can be a rule that applies a surcharge to the bill of any customer whose mobile phone minutes for the past month exceeded the allowed mobile phone minutes for the customer's service plan.

5 Execution data 56 generated by the data processing application 52 can provide information about the execution of the data processing application 52, such as rules 54 in the application 52 that did or did not execute, a number of times each rule 54 in the application 52 was executed, or other information. The execution data 56 can be provided to the data generation module 150. Based on the execution data 56, a results engine 40 in
10 the data generation module 150 (FIG. 2) can identify desired input data or input data criteria for the data processing application 52, such as input data that would have caused the unexecuted rules 54 to be executed, input data that would have caused a particular rule 54 to be executed a specified number of times, or input data that would have caused another desired execution result. For instance, the desired execution result can be
15 specified by a user, e.g., through the user interface 30.

 The data generation engine 20 of the data generation module 150 can generate data 58 to satisfy the desired input data criteria. The generated data 58 can be provided as input data to the data processing application 52. That is, for example, the data generation engine 20 can generate data to cause a previously unexecuted rule in the data processing
20 application to execute, to cause a particular rule to be executed more or fewer times, or to cause another desired execution result. In one example, a record having a value of greater than \$1,00,000 for an *income* field can be generated to cause execution of a rule that executes only for input data having *income* > \$1,000,000. In one example, a first rule executes if a *gender* field has a value of "F" and a second rule executes if the *gender* field
25 has a value of "M." If, upon initial execution of the data processing application, the execution data 56 indicate that the first rule is executed only once while the second rule is executed fifty times, data can be generated by the data generation engine 20 to satisfy the desired execution result, e.g., as specified by a user, that the first rule and the second rule are executed approximately the same number of times. For instance, the data generation
30 engine 20 can create additional records having *gender* = "F" or modify the value for the *gender* field for some of the existing records.

Referring to FIG. 20, a general approach 750 implemented by the data generation module 150 generates a set of test data that is suited for a data processing application. For instance, the set of test data can be processed by the data processing application to yield a desired execution result, such as the execution of each rule of the data processing application a desired number of times. In the approach 750, one or more records are provided for processing by a data processing application that implements one or more rules (752). Each record can have one or more fields. The data processing application can be executed (754).

Based on a number of times each rule was executed by the data processing application, a content criterion can be determined for a particular field of the one or more records by the results engine 40 of the data generation module 150 (756). In some examples, the content criterion for the particular field can be determined so that data generated according to the content criterion can achieve a desired execution result when processed by the data processing application, such as to cause the execution of a previously unexecuted rule or to cause a particular rule to be executed more or fewer times. The content criterion can be, e.g., a data characteristic for the data to be generated for the field. For instance, data characteristics for a numerical field, such as a decimal field or an integer field, can indicate a maximum allowed value, a minimum allowed value, an allowed range of values, or an average value for the field across all records. Data characteristics for a string field can indicate a maximum allowed number of characters, a minimum allowed number of characters, an average number of characters, or a set of allowed characters. Data characteristics for a temporal field, such as a date field or a datetime field, can indicate a date or datetime metric, a date or datetime basis, or a date or datetime range.

Content can be generated for the particular field based on the content criterion and based on a format specification for the particular field (758). In some examples, content for the particular field in existing records can be modified. In some examples, new records can be created and content can be generated for the particular field in the new records.

The data processing application can be executed using at least the newly generated content as input data (760). In some examples, the existing records including

modified values for the particular field can be provided to the data processing application. In some examples, the existing records can be provided to the data processing application along with newly created records. In some examples, only the newly created records can be provided to the data processing application.

5 The results of the execution of the data processing application using at least the newly generated content as input data can be analyzed (762). If the desired execution result was achieved (764), such as the rules in the data processing application being executed a desired number of times, the data generation process is completed (766). If the desired execution result was not achieved (764), further content criteria can be
10 determined (757) and new content can again be generated (758) to be used as input data for the data processing application.

 In some examples, the approach 750 can be implemented to generate a set of test data for the data processing application that causes all of the rules of the data processing application to execute at least once. That is, in order to thoroughly test the data
15 processing application, data can be provided that triggers every possible action in the data processing application. If such data do not exist, the data can be generated. For instance, once the data generation process is completed (766), the generated set of data can be processed by the data processing application to test the execution of the data processing application (768). Because the generated set of data causes every possible action in the
20 data processing application to be executed, the data processing application can be thoroughly tested. If the data processing application does not perform as expected (e.g., if an action of the data processing application does not perform as expected), the data processing application can be modified as appropriate (770) and tested until satisfactory results are achieved.

25 In some examples, the approach 750 can be implemented to generate a set of realistic data for processing by the data processing application. For instance, a data processing application for a retail corporation may generate trend reports indicative of customer buying patterns. To preserve customer privacy, a set of realistic but fictional data records can be generated for processing by the data processing application. In some
30 cases, the generated data can share characteristics with actual customer data, such as distributions of values across records of the data set (e.g., the income distribution of the

customers), without revealing sensitive personal information. In some examples, the set of realistic data can be data indicative of physical processes, data indicative of financial transactions, data indicative of human or animal behaviors, or other data.

Embodiment 1 is directed to a method including, based on a number of times each
5 of one or more rules was executed by a data processing application during processing of one or more records, determining a content criterion for each of one or more fields; generating content for each of the fields based on the content criterion; and populating each of the fields in one or more records with the generated content.

Embodiment 2 is directed to embodiment 1, wherein generating content includes
10 generating content based on a format specification for each of the fields.

Embodiment 3 is directed to embodiment 1 or 2, wherein the format specification indicates a type of content.

Embodiment 4 is directed to embodiment 3, wherein the type of content includes one or more of a string, a decimal, an integer, a date, or a datetime.

15 Embodiment 5 is directed to any of embodiments 2 to 4, including receiving the format specification through a user interface.

Embodiment 6 is directed to any of the preceding embodiments, including creating a new record including one or more of the fields; and populating the fields for the new record.

20 Embodiment 7 is directed to embodiment any of the preceding embodiments, wherein populating each of the fields with the generated content includes modifying pre-existing content for one or more of the fields.

Embodiment 8 is directed to any of the preceding embodiments, wherein populating each of the fields includes creating the particular field for one or more pre-existing records.
25

Embodiment 9 is directed to any of the preceding embodiment, wherein the content criterion includes one or more of a target value, a target range of values, a target distribution of values, a maximum value, or a minimum value.

Embodiment 10 is directed to any of the preceding embodiment, wherein the
30 content criterion includes particular values for the generated content.

Embodiment 11 is directed to any of the preceding embodiment, wherein the content criterion is indicative of a target combination of content to be included in two or more of the one or more fields.

Embodiment 12 is directed to any of the preceding embodiment, including
5 receiving the content criterion through a user interface.

Embodiment 13 is directed to any of the preceding embodiment, wherein generating content includes generating content based on a generation mode.

Embodiment 14 is directed to embodiment 13, wherein the generation mode includes one or more of (i) generation of random content for one or more of the fields for
10 each of multiple records, (ii) generation of unique content for one or more of the fields for each of multiple records, and (iii) selection of content for one or more of the fields for each of multiple records from a set of available content.

Embodiment 15 is directed to any of the preceding embodiment, including providing one or more records including the generated content to the data processing
15 application.

Embodiment 16 is directed to any of the preceding embodiment, wherein the content criterion is determined to cause a particular rule to be executed by the data processing application.

Embodiment 17 is directed to any of the preceding embodiment, wherein
20 determining the content criterion includes identifying a particular rule that was executed less than a minimum threshold number of times; and determining the content criterion to cause the particular rule to be executed at least the minimum threshold number of times.

Embodiment 18 is directed to a non-transitory computer readable medium storing instructions for causing a computer system to based on a number of times each of one or
25 more rules was executed by a data processing application during processing of one or more records, determine a content criterion for each of one or more fields; generate content for each of the fields based on the content criterion; and populate each of the fields in one or more records with the generated content.

Embodiment 19 is directed to a computing system including at least one processor
30 configured to based on a number of times each of one or more rules was executed by a data processing application during processing of one or more records, determine a

content criterion for each of one or more fields; generate content for each of the fields based on the content criterion; and populate each of the fields in one or more records with the generated content.

Embodiment 20 is directed to a computing system including means for, based on
5 a number of times each of one or more rules was executed by a data processing application during processing of one or more records, determining a content criterion for each of one or more fields; means for generating content for each of the fields based on the content criterion; and means for populating each of the fields in one or more records with the generated content.

10 Embodiment 21 is directed to a method for testing a data processing application during development of the data processing application, the method including processing first data records using the data processing application, the data processing application implementing one or more rules, each first data record having multiple fields; comparing a number of times each rule of the data processing application is executed during
15 processing of the first data records with a target value for the rule; based on the results of the comparing, determining a content criterion indicative of a target characteristic for data in second data records to be processed using the data processing application; generating content for one or more fields of the second data records according to a format of the first data records and according to the content criterion; processing the second data
20 records using the data processing application; comparing a number of times each rule of the data processing application is executed during processing of the second data records with the target value for the rule and determining that the target value for each rule is satisfied; and testing the data processing application using the second data records.

Embodiment 22 is directed to embodiment 21, including receiving the first data
25 records at a client device, wherein the data processing application is executed on the client device.

Embodiment 23 is directed to embodiment 22, wherein the client device comprises a computing device.

Embodiment 24 is directed to embodiment 21 or 22, including receiving the first
30 data records from a remote server.

Embodiment 25 is directed to embodiment 24, including receiving the first data records at a client device that is communicatively coupled to a server via a LAN or WAN connection.

Embodiment 26 is directed to any of embodiments 21 to 25, including
5 determining the content criterion such that the data in the second data records causes the target value for each rule to be satisfied when the data processing application processes the second data records.

Embodiment 27 is directed to any of embodiments 21 to 26, wherein generating content for one or more fields of the second data records includes modifying content in
10 one or more fields of the first data records, and wherein the modified first data records are the second data records.

Embodiment 28 is directed to any of embodiments 21 to 27, wherein generating content for one or more fields of the second data records includes creating a new data record.

15 Embodiment 29 is directed to any of embodiments 21 to 28, including storing the second data records in a storage device.

Embodiment 30 is directed to any of embodiments 21 to 29, including modifying the data processing application based on a result of the testing.

Embodiment 31 is directed to any of embodiments 21 to 30, wherein testing the
20 data processing application includes modifying the data processing application based on a result of processing the second data records using the data processing application.

Embodiment 32 is directed to any of embodiments 21 to 31, including determining the content criterion such that the data in the second data records is substantially similar to data expected to be processed by the data processing application
25 in a production environment.

The data generation techniques described above can be implemented using a computing system executing suitable software. For instance, the computing system can include the system 100, the development environment 118, the data generation module 150, and other computing systems. For example, the software may include procedures in
30 one or more computer programs that execute on one or more programmed or programmable computing system (which may be of various architectures such as

distributed, client/server, or grid) each including at least one processor, at least one data storage system (including volatile and/or non-volatile memory and/or storage elements), at least one user interface (for receiving input using at least one input device or port, and for providing output using at least one output device or port). The software may include
5 one or more modules of a larger program, for example, that provides services related to the design, configuration, and execution of dataflow graphs. The modules of the program (e.g., elements of a dataflow graph) can be implemented as data structures or other organized data conforming to a data model stored in a data repository.

The software may be provided on a tangible, non-transitory medium, such as a
10 CD-ROM or other computer-readable medium (e.g., readable by a general or special purpose computing system or device), or delivered (e.g., encoded in a propagated signal) over a communication medium of a network to a tangible, non-transitory medium of a computing system where it is executed. Some or all of the processing may be performed on a special purpose computer, or using special-purpose hardware, such as coprocessors
15 or field-programmable gate arrays (FPGAs) or dedicated, application-specific integrated circuits (ASICs). The processing may be implemented in a distributed manner in which different parts of the computation specified by the software are performed by different computing elements. Each such computer program is preferably stored on or downloaded to a computer-readable storage medium (e.g., solid state memory or media,
20 or magnetic or optical media) of a storage device accessible by a general or special purpose programmable computer, for configuring and operating the computer when the storage device medium is read by the computer to perform the processing described herein. The inventive system may also be considered to be implemented as a tangible, non-transitory medium, configured with a computer program, where the medium so
25 configured causes a computer to operate in a specific and predefined manner to perform one or more of the processing steps described herein.

A number of embodiments of the invention have been described. Nevertheless, is to be understood that the foregoing description is intended to illustrate and not to limit the scope of the invention, which is defined by the scope of the following claims.

30 Accordingly, other embodiments are also within the scope of the following claims. For example, various modifications may be made without departing from the scope of the

invention. Additionally, some of the steps described above may be order independent, and thus can be performed in an order different from that described.

CLAIMS

1. A computer-implemented method for generating content for data records to cause execution of a processing rule during processing of the data records, the method including:
 - processing a first set of data records by a data processing application that includes a processing rule, wherein a processing rule operates on at least one input value and generates at least one output value, and wherein whether the processing rule is executed by the data processing application during processing of a particular data record depends directly or indirectly on a value in each of one or more data fields of the particular data record;
 - receiving execution information indicative of a first number of times the processing rule was executed by the data processing application during processing of the first set of records;
 - when the first number of times the processing rule was executed is less than a threshold number of times, determining a content criterion characterizing content to be generated for each of one or more fields;
 - generating content for each of one or more fields of a second set of data records according to the content criterion; and
 - processing the second set of data records by the data processing application, wherein when the second set of data records is processed by the data processing application, the processing rule is executed by the data processing application a second number of times, and wherein the second number of times is equal to or greater than the threshold number of times.
2. The method of claim 1, wherein generating content includes generating content based on a format specification for each of the fields.
3. The method of claim 2, wherein the format specification indicates a type of content.

4. The method of claim 3, wherein the type of content includes one or more of a string, a decimal, an integer, a date, or a datetime.
5. The method of any of the claims 2 to 4, including receiving the format specification through a user interface.
6. The method of any of the preceding claims, including:
creating a new data record including one or more of the fields; and
populating the one or more fields for the new data record.
7. The method of any of the preceding claims, wherein populating the one or more data fields with the generated content includes modifying pre-existing content for one or more of the data fields of the second set of data records.
8. The method of any of the preceding claims, wherein populating the one or more of the data fields includes creating a data field for one or more pre-existing records.
9. The method of any of the preceding claims, wherein the content criterion includes one or more of a target value, a target range of values, a target distribution of values, a maximum value, or a minimum value.
10. The method of any of the preceding claims, wherein the content criterion includes particular values for the generated content.
11. The method of any of the preceding claims, wherein the content criterion is indicative of a target combination of content to be included in two or more of the one or more data fields.
12. The method of any of the preceding claims, including receiving the content criterion through a user interface.

13. The method of any of the preceding claims, wherein generating content includes generating content based on a generation mode.

14. The method of claim 13, wherein the generation mode includes one or more of (i) generation of random content for one or more of the data fields for each of multiple data records, (ii) generation of unique content for one or more of the data fields for each of multiple data records, and (iii) selection of content for one or more of the data fields for each of multiple data records from a set of available content.

15. The method of any of the preceding claims, including providing the second set of data records including the generated content to the data processing application.

16. The method of any of the preceding claims, wherein determining the content criterion includes determining the content criterion to cause a particular rule to be executed at least the target number of times.

17. A non-transitory computer readable medium storing instructions for causing a computer system to carry out the method of any of the preceding claims.

18. A computing system including at least one processor configured to execute the method of any of claims 2 to 16.

19. A computing system including:

means for processing a first set of data records by a data processing application that includes a processing rule, wherein a processing rule operates on at least one input value and generates at least one output value, and wherein whether the processing rule is executed by the data processing application during processing of a particular data record depends directly or indirectly on a value in each of one or more data fields of the particular data record;

means for receiving execution information indicative of a first number of times the processing rule was executed by the data processing application during processing of the first set of records;

means for when the first number of times the processing rule was executed is less than a threshold number of times, determining a content criterion characterizing content to be generated for each of one or more fields;

means for generating content for each of one or more fields of a second set of data records according to the content criterion; and

means for processing the second set of data records by the data processing application, wherein when the second set of data records is processed by the data processing application, the processing rule is executed by the data processing application a second number of times, and wherein the second number of times is equal to or greater than the threshold number of times.

20. A computer-implemented method for testing a data processing application during development of the data processing application, the method including:

processing first data records using the data processing application, the data processing application implementing one or more rules, each first data record having multiple fields, wherein a rule operates on at least one input value and generates at least one output value, and wherein whether a given rule is executed by the data processing application during processing of a particular data record depends directly or indirectly on a value in each of one or more of the multiple fields of the particular data record;

during or after the processing of the first data records using the data processing application, receiving execution information indicative of a number of times each of the one or more rules was executed during the processing of the first data records;

for each rule of the data processing application, comparing the number of times the rule was executed during the processing of the first data records with a target number of times for the rule;

when the number of times a given rule was executed is less than the target number of times for the rule , determining a content criterion indicative of a target characteristic for data, wherein a data record having data with the target characteristic will cause the given rule to be executed when the data record is processed using the data processing application;

generating content according to a format of the first data records and according to the content criterion;

populating one or more fields of second data records with the generated content;

processing the second data records including the generated content using the data processing application;

during or after processing of the second data records using the data processing application, receiving execution information indicative of a number of times each of the one or more rules was executed during the processing of the second data records;

for each rule of the data processing application, comparing the number of times the rule was executed during the processing of the second data records with the target number of times for the rule;

based on the comparing, determining that each rule of the data processing application was executed at least the target number of times for the rule during the processing of the second data records; and

testing the data processing application using the second data records including the generated content.

21. The method of claim 20, including receiving the first data records at a client device, wherein the data processing application is executed on the client device.

22. The method of claim 21, wherein the client device comprises a computing device.
23. The method of claim 20 or 21, including receiving the first data records from a remote server.
24. The method of claim 23, including receiving the first data records at a client device that is communicatively coupled to the remote server via a LAN or WAN connection.
25. The method of any of claims 20 to 24, wherein generating content includes modifying content in one or more fields of the first data records, and wherein the modified first data records are the second data records.
26. The method of any of claims 20 to 25, wherein generating content for one or more fields of the second data records includes creating a new data record.
27. The method of any of claims 20 to 26, including storing the second data records in a storage device.
28. The method of any of claims 20 to 27, including modifying the data processing application based on a result of the testing.
29. The method of any of claims 20 to 28, wherein testing the data processing application includes modifying the data processing application based on a result of processing the second data records using the data processing application.
30. The method of any of claims 20 to 29, including determining the content criterion such that the data in the second data records is substantially similar to data expected to be processed by the data processing application in a production environment.

31. A computer-implemented method for generating content for data records to cause execution of a processing rule during processing of the data records, the method including:

receiving execution information indicative of a first number of times a processing rule of a data processing application was executed during processing of a first set of data records by the data processing application, wherein whether the processing rule is executed by the data processing application during processing of a particular data record depends directly or indirectly on a value in each of one or more data fields of the particular data record;

when the first number of times the processing rule was executed is less than a threshold number of times, determining a content criterion indicative of a characteristic of content to be generated for each of one or more fields;

generating content for each of one or more fields of a second set of data records according to the content criterion; and

populating one or more data fields of a second set of data records with the content generated according to the content criterion, wherein the second set of data records populated with the generated content is sufficient to cause execution of the processing rule a second number of times equal to or greater than the threshold number of times when the second set of data records populated with the generated content is processed by the data processing application.

32. The method of claim 31, in which generating content includes generating content based on a format specification for each of the fields.

33. The method of any of claims 31 to 32, in which populating the one or more data fields of the second set of data records includes:

creating a new data record including the one or more fields; and

populating the one or more fields for the new data record.

34. The method of any of claims 31 to 33, in which populating the one or more data fields with the generated content includes modifying pre-existing content for one or more of the data fields of the second set of data records.
35. The method of any of claims 31 to 34, in which populating the one or more data fields includes creating a data field for one or more pre-existing data records.
36. The method of any of claims 31 to 35, in which the content criterion includes one or more of a target value, a target range of values, a target distribution of values, a maximum value, or a minimum value.
37. The method of any of claims 31 to 36, in which the content criterion includes particular values for the generated content.
38. The method of any of claims 31 to 37, in which the content criterion is indicative of a target combination of content to be included in two or more of the one or more data fields.
39. The method of one of claims 31 to 38, in which generating content includes generating content based on a generation mode.
40. The method of claim 39, in which the generation mode includes one or more of (i) generation of random content for one or more of the data fields for each of multiple data records, (ii) generation of unique content for one or more of the data fields for each of multiple data records, and (iii) selection of content for one or more of the data fields for each of multiple data records from a set of available content.
41. The method of any of claims 31 to 40, including providing the second set of data records including the generated content to the data processing application.

42. The method of any of claims 31 to 41, in which determining the content criterion includes determining the content criterion to cause the processing rule to be executed a number of times equal to or greater than the threshold number of times.

43. A non-transitory computer readable storage medium storing instructions for causing a computing system to carry out the method of any of claims 31 to 42.

44. A computing system including at least one processor configured to execute the method of any of claims 32 to 42:

45. A computing system for generating content for data records to cause execution of a processing rule during processing of the data records, the computing system including:

means for receiving execution information indicative of a first number of times a processing rule of a data processing application was executed during processing of a first set of data records by the data processing application, wherein whether the processing rule is executed by the data processing application during processing of a particular data record depends directly or indirectly on a value in each of one or more data fields of the particular data record;

means for, when the first number of times the processing rule was executed is less than a threshold number of times, determining a content criterion indicative of a characteristic of content to be generated for each of one or more fields;

generating content for each of one or more fields of a second set of data records according to the content criterion; and

means for populating one or more data fields of a second set of data records with the content generated according to the content criterion, wherein the second set of data records populated with the generated content is sufficient to cause execution of the processing rule a second number of times equal to or greater than the threshold number of times when the second set of data records populated with the generated content is processed by the data processing application.