

WHAT IS CLAIMED IS:

1. A method for determining a data quality rule for values in a field of a data record in a set of data records based on a label associated with the field, the label indicating characteristics of the values of the field, the method being implemented by a data processing system and including:

retrieving a label index that associates a label with a set of one or more fields in a data record, wherein the label identifies a type of information expected in each field of the set of one or more fields;

accessing a data dictionary that associates the type of information indicated by the label with a set of attribute values representing requirements for values of the one or more fields associated with the label, the requirements including logical or syntactical characteristics of the values for the one or more fields; and

for a field of a particular data record:

identifying, by accessing the label index, a particular label associated with the field of the particular data record;

retrieving, from the data dictionary, an attribute value for the particular label, the attribute value specifying a particular requirement for the field; and

generating a data quality rule that, when executed, is configured to:

validate whether a value of the field satisfies the particular requirement represented by the attribute value, and

generate output data indicating whether the particular requirement is satisfied.

2. The method of claim 1, wherein the data quality rule is indicative of one or more of:

an allowable deviation for the value in the field from the requirement,

one or more allowable values in the field, and

one or more prohibited values in the field.

3. The method of claim 2, wherein the one or more allowable values or prohibited values in the field are associated with a field name of the field.

4. The method of claim 2 or 3, wherein the field is a first field, and wherein the one or more allowable values or prohibited values in the field are determined based on a value in a second field in the particular data record, the second field being related to the first field of the particular data record.

5. The method of any of claims 2 to 4, wherein the one or more allowable values or prohibited values are based on a combination of the value in the second field and the value in the first field.

6. The method of any of claims 2 to 5, wherein the one or more allowable values correspond to values that satisfy a numerical function, and wherein the one or more prohibited values correspond to values that do not satisfy the numerical function.

7. The method of and of claims 1 to 6, wherein the field is a first field, wherein the data record comprises a second field, wherein the data quality rule is a first data quality rule, and wherein the method comprises:

determining, based on the attribute value for the particular label, that a relationship exists between the first field and the second field; and

generating a second data quality rule for the second field based on the relationship between the second field and the first field.

8. The method of claim 7, wherein the relationship is indicative of a dependency of a value of the second field on a value of the first field or the value of the first field on the value of the second field.

9. The method of claim 7 or 8, wherein the relationship is indicative of a correlation between a value of the first field and a value of the second field.

10. The method of any of claims 7 to 9, further comprising obtaining validation data that validates the relationship for each value of the first field and the second field.

11. The method of any of claims 7 to 10, wherein the relationship is validated for a threshold number of values for the first field and the second field.

12. The method of any of claims 7 to 11, wherein the second data quality rule is configured to enforce a constraint on a value of the second field based on a value of the first field.

13. The method of any of claims 7 to 12, wherein determining that the relationship exists between the first field and the second field comprises determining that a value of the second field comprises a key value referenced by a value of the first field, and wherein the data quality rule is configured to require that each value of the second field is a valid key value.

14. The method of any of claims 7 to 13, wherein the first field and the second field each comprises numeric values, and wherein determining that the relationship exists between the first field and the second field comprises determining a numerical function that relates values of the first field to values of the second field.

15. The method of any of claims 7 to 14, wherein determining that the relationship exists between the first field and the second field comprises using at least one classifier configured by a machine learning process.

16. The method of any of claims 1 to 15, comprising:
determining that the attribute value associated with the particular label indicates that the field comprises primary key values for the particular data record;
and
configuring the data quality rule to require that the primary key values are each unique in the field.

17. The method of any of claims 1 to 16, the attribute value for the particular label represents at least one of an average for the values in the field, a maximum length for the values, a minimum length for the values, a data type for the values, and a format for the values.

18. The method of any of claims 1 to 17, wherein the particular data record is a first data record, and wherein the method comprises applying the data quality rule to another field associated with the particular label in a second data record that is different than the first data record.

19. The method of any of claims 1 to 18, wherein the method comprises storing data associating the data quality rule and the particular label.

20. The method of any of claims 1 to 19, further comprising:
generating data for requesting approval of the data quality rule; and
approving the data quality rule in response to obtaining approval data indicative of approval of the data quality rule.

21. The method of any of claims 1 to 20, wherein generating the data quality rule includes:
determining a historical trend for values in the field; and
generating a requirement based on the historical trend for the values in the field.

22. The method of any of claims 1 to 21, wherein generating the data quality rule for the field includes:
identifying a historical average of a value in the field; and
generating a requirement based on the historical average of the value in the field.

23. A method for selecting test data to cause execution of a processing rule during testing of a data processing application, the method including:
retrieving a label index that associates a label with a set of one or more fields in a data record, wherein the label identifies a type of information expected in each field of the set of one or more fields;
accessing a data dictionary that associates the type of information indicated by the label with a set of attribute values representing requirements for values of the one

or more fields associated with the label, the requirements including logical or syntactical characteristics of the values for the one or more fields; and

for a field of a particular data record:

identifying, by accessing the label index, a particular label associated with the field of the particular data record;

retrieving, from the data dictionary, an attribute value for the particular label, the attribute value specifying a particular requirement for the field;

generating a subsetting rule that, when executed, specifies whether a value of the field includes the attribute value;

selecting a subset of fields from the particular data record according to the subsetting rule, wherein each selected field includes values that have the attribute value; and

providing the selected subset to a data processing application for testing of the data processing application.

24. A method for masking data of a data processing application, the method comprising:

retrieving a label index that associates a label with a set of one or more fields in a data record, wherein the label identifies a type of information expected in each field of the set of one or more fields;

accessing a data dictionary that associates the type of information indicated by the label with a set of attribute values representing requirements for values of the one or more fields associated with the label, the requirements including logical or syntactical characteristics of the values for the one or more fields; and

for a field of a particular data record:

identifying, by accessing the label index, a particular label associated with the field of the particular data record;

retrieving, from the data dictionary, an attribute value for the particular label, the attribute value specifying a particular requirement for the field; and

determining, based on the attribute value, that the field of the particular data record represents sensitive data; and

in response to the determining, performing a data masking function to transform values including the sensitive data of the field into masked values.

25. A method for determining a schema of a data record, the method comprising:

- retrieving a label index that associates a label with a set of one or more fields in a data record, wherein the label identifies a type of information expected in each field of the set of one or more fields;

- accessing a data dictionary that associates a type of information indicated by the label with a set of attribute values representing requirements for values of the one or more fields associated with the label, the requirements including logical or syntactical characteristics of the values for the one or more fields; and

- for a first field of a particular data record:

 - identifying, by accessing the label index, a particular label associated with the first field of the particular data record;

 - retrieving, from the data dictionary, an attribute value for the particular label, the attribute value specifying a particular requirement for the first field;

 - determining that the attribute value represents a schema feature for values included in the first field;

 - determining, based on the schema feature, that the first field having the label comprises key values being referenced by values of a second field; and

 - in response to the determining, updating schema data describing the particular data record to reference that the first field comprises the key values being referenced by the values of the second field.

26. A method for determining a data quality rule for values in one or more datasets, the method being implemented by a data processing system and comprising:

- profiling, by the data processing system, values stored in a field included in data records of one or more datasets;

- applying, by the data processing system, one or more classifiers to the profiled values;

- identifying, based on applying the one or more classifiers, one or more attributes indicative of a logical or syntactical characteristic for the values of the field, with each of the one or more attributes having a respective confidence level that is based on an output of each of the one or more classifiers;

associating, with the field, one or more of the identified attributes for which the confidence level satisfies a threshold level;

determining, based on the one or more attributes associated with the field, one or more constraints for values included in the field; and

based on the determined one or more constraints, determining a data quality rule for the field of the dataset, the data quality rule indicating a relationship between values of the field and the one or more constraints.

27. A data processing system for determining a data quality rule for values in a field of a data record in a set of data records based on a label associated with the field, the label indicating characteristics of the values of the field, the data processing system including:

at least one processing device; and

at least one memory in communication with the at least one processing device, the at least one memory storing instructions that, when executed by the at least one processing device, cause the at least one processing device to perform operations including:

retrieving a label index that associates a label with a set of one or more fields in a data record, wherein the label identifies a type of information expected in each field of the set of one or more fields;

accessing a data dictionary that associates the type of information indicated by the label with a set of attribute values representing requirements for values of the one or more fields associated with the label, the requirements including logical or syntactical characteristics of the values for the one or more fields; and

for a field of a particular data record:

identifying, by accessing the label index, a particular label associated with the field of the particular data record;

retrieving, from the data dictionary, an attribute value for the particular label, the attribute value specifying a particular requirement for the field; and

generating a data quality rule that, when executed, is configured to:

validate whether a value of the field satisfies the particular requirement represented by the attribute value, and generate output data indicating whether the particular requirement is satisfied.

28. One or more non-transitory computer readable media storing instructions for determining a data quality rule for values in a field of a data record in a set of data records based on a label associated with the field, the label indicating characteristics of the values of the field, the instructions being configured, when executed by at least one processing device, to cause the at least one processing device to perform operations including:

retrieving a label index that associates a label with a set of one or more fields in a data record, wherein the label identifies a type of information expected in each field of the set of one or more fields;

accessing a data dictionary that associates the type of information indicated by the label with a set of attribute values representing requirements for values of the one or more fields associated with the label, the requirements including logical or syntactical characteristics of the values for the one or more fields; and

for a field of a particular data record:

identifying, by accessing the label index, a particular label associated with the field of the particular data record;

retrieving, from the data dictionary, an attribute value for the particular label, the attribute value specifying a particular requirement for the field; and

generating a data quality rule that, when executed, is configured to:

validate whether a value of the field satisfies the particular requirement represented by the attribute value, and

generate output data indicating whether the particular requirement is satisfied.