

What is claimed is:

1. A method including:
 - receiving information indicative of an output dataset generated by a data processing system;
 - identifying, based on data lineage information relating to the output dataset, one or more upstream datasets on which the output dataset depends;
 - analyzing one or more of the identified one or more upstream datasets on which the output dataset depends, the analyzing including:
 - for each particular upstream dataset of the one or more upstream datasets, applying one or more of:
 - (i) a first rule indicative of an allowable deviation between a profile of the particular upstream dataset and a reference profile for the particular upstream dataset, and
 - (ii) a second rule indicative of one or more allowable values or prohibited values for each of one or more data elements in the particular upstream dataset;
 - based on the results of applying the one or more rules, selecting one or more of the upstream datasets; and
 - outputting information associated with the selected one or more upstream datasets.
2. A non-transitory computer-readable medium storing instructions for causing a computing system to:
 - receive information indicative of an output dataset generated by a data processing system;
 - identify, based on data lineage information relating to the output dataset, one or more upstream datasets on which the output dataset depends;
 - analyze one or more of the identified one or more upstream datasets on which the output dataset depends, the analyzing including:

for each particular upstream dataset of the one or more upstream datasets,
applying one or more of:

- (i) a first rule indicative of an allowable deviation between a
profile of the particular upstream dataset and a reference
profile for the particular upstream dataset, and
- (ii) a second rule indicative of one or more allowable values or
prohibited values for each of one or more data elements in
the particular upstream dataset;

based on the results of applying the one or more rules, selecting one or
more of the upstream datasets; and

output information associated with the selected one or more upstream datasets.

3. A computing system including:

a processor coupled to a memory, the processor and memory configured to:

receive information indicative of an output dataset generated by a data
processing system;

identify, based on data lineage information relating to the output dataset,
one or more upstream datasets on which the output dataset
depends;

analyze one or more of the identified one or more upstream datasets on
which the output dataset depends, the analyzing including:

for each particular upstream dataset of the one or more upstream
datasets, applying one or more of:

- (i) a first rule indicative of an allowable deviation between
a profile of the particular upstream dataset and a
reference profile for the particular upstream dataset,
and
- (ii) a second rule indicative of one or more allowable
values or prohibited values for each of one or more
data elements in the particular upstream dataset;

based on the results of applying the one or more rules, selecting one or more of the upstream datasets; and
output information associated with the selected one or more upstream datasets.

4. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which one or more of the first rule and the second rule are automatically generated.

5. The method, or the non-transitory computer-readable medium, or the computing system of claim 4, in which the first rule is automatically generated based on an automated analysis of historical profiles of the particular upstream dataset.

6. The method, or the non-transitory computer-readable medium, or the computing system of claim 5, in which the reference profile is based on an historical average profile for the particular upstream dataset.

7. The method, or the non-transitory computer-readable medium, or the computing system of claim 4, in which the second rule is automatically generated based on an automated analysis of historical values for the one or more data elements in the particular upstream dataset.

8. The method, or the non-transitory computer-readable medium, or the computing system of claim 7, in which the allowable value or prohibited value is determined based on the automated analysis.

9. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which one or more of the first rule and the second rule are specified by a user.

10. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, including receiving a specification of one or more of the first rule and the second rule through a user interface.

11. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which data lineage information indicates one or more datasets that the output dataset depends on, one or more datasets that depend on the output dataset, or both.

12. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which analyzing each of the one or more datasets to identify a subset of the datasets includes determining which of the one or more datasets have errors or possible errors; and

including selecting the datasets that have errors or possible errors for the subset.

13. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which analyzing each of the one or more datasets to identify a subset of the datasets includes identifying a particular dataset for which the deviation between the profile of the particular dataset and the reference profile for the particular dataset exceeds the allowable deviation indicated by the corresponding first rule; and

including selecting the particular dataset for the subset.

14. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which analyzing each of the one or more datasets to identify a subset of the datasets includes identifying a particular dataset having a data element with a value that does not satisfy the allowable or prohibited value indicated by the corresponding second rule; and

including selecting the particular dataset for the subset.

15. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, including identifying a data element in the output dataset, and in which identifying the one or more datasets on which the output dataset depends includes identifying datasets that affect the identified data element in the output dataset.

16. The method, or the non-transitory computer-readable medium, or the computing system of claim 15, in which identifying a data element in the output dataset includes identifying a data element that has an error or a possible error.

17. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, including generating a profile of one or more of the upstream datasets.

18. The method, or the non-transitory computer-readable medium, or the computing system of claim 17, in which generating a profile of a particular dataset includes generating a new profile of the particular dataset when a new version of the particular dataset is received.

19. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which the reference profile for a particular dataset is derived from one or more previous profiles of the particular dataset.

20. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which outputting information associated with the subset of datasets includes outputting an identifier of each of the datasets of the subset.

21. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which outputting information associated with the subset of datasets includes outputting an indicator of an error or possible error associated with each of the datasets of the subset.

22. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, including displaying a representation of the data processing system on a user interface, and in which outputting information associated with the subset of datasets includes displaying information associated with a particular dataset of the subset in a vicinity of a representation of the particular dataset of the subset of datasets.

23. The method, or the non-transitory computer-readable medium, or the computing system of claim 22, in which the displayed information associated with the particular dataset of the subset includes a value indicative of a deviation between the profile of the particular dataset and the reference profile for the particular dataset.

24. The method, or the non-transitory computer-readable medium, or the computing system of claim 22, in which the displayed information associated with the particular dataset of the subset includes a value representative of a number of data elements in the particular dataset that do not satisfy the allowable value or prohibited indicated by the corresponding second rule.

25. The method, or the non-transitory computer-readable medium, or the computing system of claim 22, including displaying an information bubble or pop-up window showing information about the subset of datasets.

26. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, including providing a user interface to enable a user to add a rule, modify a rule, or remove a rule.

27. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which the datasets include one or more source datasets and one or more reference datasets, the source datasets including data elements to be processed by the data processing system, the reference datasets including reference values that are referenced by the data processing system in processing the data elements in the source datasets.

28. The method, or the non-transitory computer-readable medium, or the computing system of claim 27, in which the reference datasets include data associated with a business entity associated with the data processing system and the source datasets include data associated with customers of the business entity.

29. The method of claim 1, or the non-transitory computer-readable medium of claim 2, or the computing system of claim 3, in which the data processing system includes transform elements, and the method includes identifying, based on the data lineage information, one or more transform elements that affect the output dataset.

30. The method, or the non-transitory computer-readable medium, or the computing system of claim 29, including determining which one or more of the transform elements have errors or possible errors.

31. The method, or the non-transitory computer-readable medium, or the computing system of claim 30, including determining whether a particular data processing elements has errors or possible errors based on an implementation date associated with the particular transform element.

32. A computing system including:

means for receiving information indicative of an output dataset generated by a data processing system;

means for identifying, based on data lineage information relating to the output dataset, one or more upstream datasets on which the output dataset depends;

means for analyzing one or more of the identified one or more upstream datasets on which the output dataset depends, the analyzing including:

for each particular upstream dataset of the one or more upstream datasets, applying one or more of:

- (i) a first rule indicative of an allowable deviation between a profile of the particular upstream dataset and a reference profile for the particular upstream dataset, and
- (ii) a second rule indicative of one or more allowable values or prohibited values for each of one or more data elements in the particular upstream dataset;

based on the results of applying the one or more rules, selecting one or more of the upstream datasets; and

means for outputting information associated with the selected one or more upstream datasets.

33. A method including:

upon identifying an error or possible error in a data element of a downstream dataset of a data processing system, automatically identifying, based on data lineage information relating to the downstream dataset, one or more upstream datasets that affect the data element;

determining which of the upstream datasets have errors or possible errors, including analyzing a current profile and a reference profile of each of the identified upstream datasets; and

outputting information associated with each of the upstream datasets that are determined to have errors or likely have errors.