

Claims

1. A method for managing lineage information, the method including:
receiving lineage information representing one or more lineage relationships among two or more data processing programs and two or more logical datasets, wherein at least one of the logical datasets resolves to a physical dataset at run time of at least one of the data processing programs;
receiving one or more runtime artifacts, each runtime artifact including information related to a previous execution of a data processing program of the two or more data processing programs; and
analyzing the one or more runtime artifacts and the lineage information and determining one or more candidate modifications to the lineage information based on results of the analyzing, wherein at least one candidate modification includes a modification to a representation of at least one of the two or more logical datasets based at least in part on a result of the analyzing, wherein the result is associated with one or more physical datasets;
wherein the one or more candidate modifications include a first candidate modification that adds a new direct lineage relationship between a data processing program of the two or more data processing programs and a logical dataset of the two or more logical datasets;
wherein analyzing the runtime artifacts and the lineage information includes analyzing logs of previous executions of the two or more data processing programs to determine physical datasets read from or written to by the two or more data processing programs; and
wherein analyzing the runtime artifacts and the lineage information further includes identifying two distinct logical datasets of the two or more logical datasets that are represented in the lineage information and are associated with the same physical dataset.

2. The method of claim 1, wherein the one or more candidate modifications include a second candidate modification that adds a new indirect lineage relationship between a data processing program of the two or more data processing programs and a logical dataset of the two or more logical datasets.

3. The method of claim 1, wherein the first candidate modification includes creation of the new direct lineage relationship between the two distinct logical datasets.

4. The method of claim 1, wherein the first candidate modification includes creation of the new direct lineage relationship including merging the two distinct logical datasets into a new combined logical dataset.

5. The method of claim 1, wherein each data processing program of the two or more data processing programs is an instance of a generic data processing program instantiated according to a set of one or more parameter values.

6. The method of claim 5, wherein analyzing the one or more runtime artifacts and the lineage information includes:

analyzing one or more logs of previous executions of a first data processing program of the two or more data processing programs to determine a first parameter set used in a first instantiation of the first data processing program according to a first set of one or more parameter values,
selecting at least some parameters from the first parameter set, and
determining that the first instantiation of the first data processing program is not represented in the lineage information based on a generic version of the first data processing program and the at least some parameters.

7. The method of claim 6, wherein selecting at least some parameters from the first parameter set includes selecting parameters based on information received from a user.

8. The method of claim 6, wherein selecting at least some parameters from the first parameter set includes selecting parameters based on one or more predefined rules.

9. The method of claim 8, wherein a first rule of the one or more predefined rules specifies that parameters with parameter values in the form of a date are excluded from the selected parameters.

10. The method of claim 8, wherein a first rule of the one or more predefined rules specifies that a parameter with a parameter value that is transformed in the logic of a generic data processing program is included in the selected parameters.

11. The method of claim 1, wherein the results of the analyzing include identification of at least one physical dataset to which at least one logical dataset resolved at run time of at least one data processing program.

12. The method of claim 1, further including applying a selected one of the one or more candidate modifications to the lineage information.

13. The method of claim 12, wherein the selected candidate modification is selected based at least in part on user input received after presenting one or more of the candidate modifications.

14. A computer-readable storage medium comprising instructions which, when executed by a computing system, cause the computing system to carry out the method of any one of claims 1-13.

15. A computing system comprising at least one processor configured to perform the method of any one of claims 1-13.