

What is claimed is:

1. A computer-implemented method, the method including:
 - receiving data lineage information for multiple components representing at least two logical datasets and a transformation, with the data lineage information identifying a first logical dataset from which the transformation is to receive data and a second logical data dataset to which transformed data is to be provided, the transformation represented by a component in the data lineage information including one or more rules to be applied to data from the first logical dataset, and with the data lineage information identifying paths and flow traces of data though the multiple components;
 - receiving data specifying one or more proposed changes to a field in the first logical dataset, a field in the second logical dataset, or the transformation;
 - analyzing the multiple components according to the data lineage information to identify each component affected by the one or more proposed changes;
 - for at least one identified component identified as affected by the one or more proposed changes,
 - generating an impact metric representing a number of times a field specified by the one or more proposed changes is referenced within the at least one identified component;
 - determining, based on the generated impact metric, an impact of implementing the one or more proposed changes to one or more of the multiple components; and
 - storing information about the impact metric.
2. The method of claim 1, wherein the impact metric provides a measure of a direct impact of implementing the one or more proposed changes to the one or more of the multiple components.
3. The method of claim 1, wherein the impact metric provides a measure of an indirect impact of implementing the one or more proposed changes to the one or more of the multiple components.

4. The method of claim 1, further comprising associating a cost to a proposed change based on one or more impact metrics.
5. The method of claim 1, further including:
 - for that at least one component identified that represents the transformation, generating annotated code that includes the impact metric;
 - compiling the annotated code into an executable program; and
 - at least partially based on execution of the executable program, determining the impact of the one or more proposed changes to one or more of the multiple components.
6. The method of claim 1, further including:
 - selecting from data storage stored items of annotated code for one or more of the multiple components in the data lineage information, which are affected by the one or more proposed changes;
 - compiling the annotated code of one or more of the stored items into a transform;
 - and
 - processing the transform to determine a risk output of a corresponding impact metric.
7. A system comprising:
 - one or more computers and one or more storage devices storing instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform operations comprising:
 - receiving a data lineage information for multiple components representing at least two logical datasets and a transformation, with the data lineage information identifying a first logical dataset describing characteristics of a first physical dataset from which the transformation is to receive data and a second logical dataset describing characteristics of a second physical dataset to which transformed data is to be provided, the transformation represented by a component in the data lineage information including one or more rules to be

applied to data from the first logical dataset, and with the data lineage information identifying paths and flow traces of data through the multiple components;

receiving data specifying one or more proposed changes to a field in the first logical dataset, a field in the second logical dataset, or the transformation;

analyzing the multiple components according to the data lineage information to identify each component affected by the one or more proposed changes;

for at least one component identified as affected by the one or more proposed changes,

generating an impact metric representing a number of times a field specified by the one or more proposed changes is referenced within the at least one identified component;

determining, based on the generated impact metric, an impact of implementing the one or more proposed changes to one or more of the multiple components; and

storing information about the impact metric.

8. The system of claim 7, wherein the impact metric provides a measure of a direct impact of implementing the one or more proposed changes to the one or more of the multiple components.

9. The system of claim 7, wherein the impact metric provides a measure of an indirect impact of implementing the one or more proposed changes to the one or more of the multiple components.

10. The system of claim 7, wherein a proposed change is one of the group consisting of a change in format of a field in a dataset or a change in a transformation.

11. The system of claim 7, wherein the operations further comprise associating a cost to a proposed change based on one or more impact metrics.

12. A system comprising:
one or more processor devices; and
memory operatively coupled to the one or more processor devices, storing a computer program that configures the system to:
- receive data lineage information for multiple components representing at least first and second logical datasets and a transformation, with the data lineage information identifying a first physical dataset from which the transformation is to receive data and a second physical data dataset to which transformed data is to be provided, the transformation represented by a component in the data lineage information including one or more rules to be applied to data from the first logical dataset, and with the data lineage information identifying paths and flow traces of data through the multiple components;
 - receive data specifying one or more proposed changes to a field in the first logical dataset, the second logical dataset, or the transformation;
 - analyze the multiple components according to the data lineage information to identify each component affected by the one or more proposed changes;
 - for at least one identified component identified as affected by the one or more proposed changes,
 - generate an impact metric representing a number of times a field specified by the one or more proposed changes is referenced within the at least one identified component;
 - determine, based on the generated impact metric, an impact of implementing one or more of the multiple components; and
 - store information about the impact metric.

13. A computer storage medium encoded with computer program instructions that when executed by one or more computers cause the one or more computers to perform operations comprising:
- receiving data lineage information for multiple components representing at least two logical datasets and a transformation, with the data lineage information identifying a first logical dataset describing characteristics of a first physical dataset from which the

transformation is to receive data and a second logical dataset describing characteristics of a second physical dataset to which transformed data is to be provided, the transformation represented by a component in the data lineage information including one or more rules to be applied to data from the first logical dataset, and with the data lineage information identifying paths and flow traces of data through the multiple components;

receiving data specifying one or more proposed changes to a field in the first logical dataset, a field in the second logical dataset, or the transformation;

analyzing the multiple components according to the data lineage information to identify each component affected by the one or more proposed changes;

for at least one component identified that represents a transformation,

for at least one identified component identified as affected by the one or more proposed changes,

generating an impact metric representing a number of times a field specified by the one or more proposed changes is referenced within the at least one identified component;

determining, based on the generated impact metric, an impact of implementing the one or more proposed changes to one or more of the multiple components; and

storing information about the impact metric.

14. The medium of claim 13, where the impact metric provides a measure of a direct impact of implementing the one or more proposed changes to the one or more of the multiple components.

15. The medium of claim 13, where the impact metric provides a measure an indirect impact of implementing the one or more proposed changes to the one or more of the multiple components.

16. The medium of claim 13, wherein a proposed change is one of the group consisting of a change in format of a field in a dataset or a change in a transformation.

17. The medium of claim 13, further comprising associating a cost to a proposed change based on one or more impact metrics.