

What is claimed is:

1. A method for managing lineage information in a computing system, the method including:

storing, in a data storage system, a specification of a directed graph that includes a plurality of nodes and directed links that represent respective lineage relationships between the plurality of nodes, at least some of the nodes being associated with one or more tag identifiers of a plurality of tag identifiers, at least one tag identifier of the plurality of tag identifiers having a plurality of possible tag values;

processing, using at least one processor, a request for a representation of data lineage for a first node of the plurality of nodes, the processing including:

determining an association between the first node and at least a first tag identifier of the plurality of tag identifiers, and determining a first subset of at least one and fewer than all of a plurality of possible tag values for the first tag identifier; and

traversing nodes along a first lineage path of directed links from the first node to determine a data lineage for the first node, where the data lineage includes one or both of (1) a dependency lineage that consists of a unidirectional upstream lineage path of directed links upstream of the first node, and (2) an impact lineage that consists of a unidirectional downstream lineage path of directed links downstream of the first node, and where determining the data lineage includes, for each traversed node determining whether to add the traversed node to the data lineage or to exclude the traversed node from the data lineage based at least in part on any tag identifiers or tag values associated with the traversed node.

2. The method of claim 1 wherein determining the data lineage includes, for each traversed node:

adding the traversed node to the data lineage if the traversed node is not associated with the first tag identifier, or the traversed node is associated with the first tag identifier and is associated with at least one tag value of the first subset; or

excluding the traversed node from the data lineage if the traversed node is associated with the first tag identifier and is not associated with at least one tag value of the first subset of the plurality of possible tag values for the first tag identifier.

3. The method of claim 2 further including, upon excluding a traversed node from the data lineage, terminating traversing of the first lineage path.

4. The method of claim 1 further including analyzing the plurality of nodes, prior to traversing the nodes of the plurality of nodes, to determine a first subset of the plurality of nodes that are associated with the first tag identifier and are not associated with at least one tag value of the first subset of the plurality of possible tag values for the first tag identifier.

5. The method of claim 4 wherein determining whether to add the traversed node to the data lineage or to exclude the traversed node from the data lineage includes determining whether the traversed node is included in the first subset of the plurality of nodes.

6. The method of claim 5 wherein determining whether to add the traversed node to the data lineage includes adding the traversed node to the data lineage if the traversed node is not included in the first subset of the plurality of nodes and excluding the traversed node from the data lineage if the traversed node is included in the first subset of the plurality of nodes.

7. The method of claim 1 wherein at least some nodes of the plurality of nodes are nested within other nodes of the plurality of nodes.

8. The method of claim 7 wherein tag identifiers and their associated subsets of tag values associated with nodes at lower levels of a nested hierarchy of nodes are given priority over tag identifiers and their associated subsets of tag values associated with nodes at higher levels of the nested hierarchy of nodes.

9. The method of claim 1 wherein traversing the nodes along the first lineage path includes traversing the nodes beginning from the first node and in a direction opposite a direction of the directed links.

10. The method of claim 1 further including generating the specification of the directed graph, where one or more of the plurality of nodes represent computation, one or more of the plurality of nodes represent data elements received or produced by a computation during execution of the computation, and the directed links representing lineage relationships between a computation and a data element.

11. The method of claim 10 wherein one or more of the plurality of nodes represents a grouping of one or more other nodes of the plurality of nodes.

12. The method of claim 1 wherein the first subset of the plurality of possible tag values for the first tag identifier consists of one or more tag values that are associated with the first node.

13. The method of claim 1 wherein the first subset of the plurality of possible tag values for the first tag identifier consists of one or more tag values that are received from user input.

14. The method of claim 1 further including traversing nodes along a plurality of lineage paths of directed links from the first node to determine a data lineage for the first node, where determining the data lineage includes, for each traversed node determining whether to add the traversed node to the data lineage or to exclude the traversed node from the data lineage based at least in part on any tag identifiers or tag values associated with the traversed node.

15. The method of claim 1 further including adding the first node to the data lineage by default.

16. Software stored on a computer-readable medium, for managing lineage information in a computing system, the software including instructions for causing a computing system to perform the method of any one of claims 1 to 15.

17. A computing system for managing lineage information, the computing system including:

a data storage system storing a specification of a directed graph that includes a plurality of nodes and directed links that represent respective lineage relationships between the plurality of nodes, at least some of the nodes being associated with one or more tag identifiers of a plurality of tag identifiers, at least one tag identifier of the plurality of tag identifiers having a plurality of possible tag values; and

at least one processor configured to perform the method of any one of claims 1 to 15.

18. A computing system for managing lineage information, the computing system including means for performing the steps of the method of any one of claims 1 to 15.