What is claimed is:

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

receiving, over an input device or port, an identification of a directed data lineage graph that includes one or more data transformation nodes that represent computations performed by the computing system that receive data as input, transform the received data, and produce the transformed data as output, and one or more data nodes that represent data including the data received as input and the data produced as output, and includes directed links that represent respective data lineage relationships between a computation and data to be received or produced by the computation during execution of the computation; and

computing, using at least one processor, summary information based on paths in the directed data lineage graph, and storing the summary information in one or more summary objects, the computing including

receiving designation of interest for a plurality of the nodes of the directed data lineage graph; and

generating one or more summary objects for remaining nodes not included in the plurality of nodes of interest, a first summary object of the one or more summary objects including summary information based on a first path between a first node of interest and a second node of interest that does include one or more of the remaining nodes and does not include any nodes of interest other than the first and second nodes; wherein generating the first summary object includes traversing the first path between the first node of interest and the second node of interest to identify one or more remaining nodes that are not designated as being of interest along the first path, and forming the summary information for the identified one or more remaining nodes.

2. The method of claim 1 further including generating a data lineage diagram as a visual representation of at least portions of the directed data lineage graph, in which each node designated as being of interest is represented by its own icon, and a plurality of the remaining nodes are represented

by a common icon connected to a plurality of nodes of interest according to summary information stored in one or more of the summary objects.

3. The method of claim 2 further including receiving a user input indicative of a user's desire to view details associated with the common icon and, in response to the user input, traversing the plurality of remaining nodes represented by the common icon and visually representing the remaining nodes in the data lineage diagram based on the traversal of paths of the directed data lineage graph including the plurality of remaining nodes associated with the summary information.

4. The method of claim 2 further including receiving a designation of one of the one or more data transformation nodes or one of the one or more data nodes as a target node, wherein generating the data lineage diagram includes traversing one or more paths through the directed data lineage graph, each path of the one or more paths including two or more nodes of interest and the target node.

5. The method of claim 4 wherein for at least some of the one or more paths through the directed data lineage graph, traversing the path includes traversing a first portion of the path in an upstream direction from the target node and traversing a second portion of the path in a downstream direction from the target node.

6. The method of claim 5 wherein traversing the first portion of the path includes determining whether the target node is marked as being a node of interest, and

if the target node is determined to be a node of interest, traversing a link corresponding to a summary object associated with the target node from the target node to a first upstream node of interest, otherwise

if the target node is not determined to be a node of interest, traversing an original path of the directed data lineage graph between the target node and the first upstream node of interest, including traversing one or more data transformation nodes or data nodes that are not designated as being of interest.

7. The method of claim 6 wherein traversing the first portion of the path further includes traversing a link corresponding to a summary object associated with the first upstream node of interest from the first upstream node of interest to a second upstream node of interest.

8. The method of claim 5 wherein traversing the second portion of the path includes determining whether the target node is marked as being a node of interest, and

if the target node is determined to be a node of interest, traversing a link corresponding to a summary object associated with the target node from the target node to a first downstream node of interest, otherwise

if the target node is not determined to be a node of interest, traversing an original path of the directed data lineage graph between the target node and the first downstream node of interest, including traversing one or more data transformation nodes or data nodes that are not designated as being of interest.

9. The method of claim 8 wherein traversing the second portion of the path further includes traversing a link corresponding to a summary object associated with the first downstream node of interest from the first downstream node of interest to a second downstream node of interest.

10. The method of claim 2 wherein the one or more summary objects includes two or more summary objects, generating the lineage diagram further includes merging at least some of the two or more summary objects into a summary node represented by the common icon.

11. The method of claim 10 wherein merging at least some of the two or more summary objects into the summary node includes analyzing data lineage relationships between the nodes of interest linked by the at least some of the two or more summary objects to determine whether merging the at least some of the two or more summary objects is possible.

12. The method of claim 11 wherein analyzing data lineage relationships between the nodes of interest includes determining that the at least some of the two or more summary objects are associated with a common downstream node of interest.

13. The method of claim 1, wherein the designation of interest for at least some of the nodes is received from a user.

14. The method of claim 1, wherein the designation of interest for at least some of the nodes is generated by the computing system.

15. The method of claim 14, wherein the computing system generates the designation of interest for at least some of the nodes based on a position of the at least some of the nodes relative to a position of other nodes designated as being of interest.

16. The method of claim 14, wherein the computing system generates the designation of interest for at least some of the nodes based on the at least some of the nodes sharing a data structure with other nodes designated as being of interest.

17. The method of claim 2 wherein the computing system includes a plurality of separate, interconnected sub-systems, the portions of the directed data lineage graph being distributed among at least some of the sub-systems.

18. The method of claim 17 wherein at least some of the sub-systems are represented by the common icon.

19. Software stored in a non-transitory form on a computer-readable medium, for managing lineage information in a computing system, the software including instructions for causing a computing system to:

receive, over an input device or port, an identification of a directed data lineage graph that includes one or more data transformation nodes that represent computations performed by the computing system that receive data as input, transform the received data, and produce the transformed data as output, and one or more data nodes that represent data including the data received as input and the data produced as output, and includes directed links that represent respective data lineage relationships between a computation and data to be received or produced by the computation during execution of the computation; and

compute, using at least one processor, summary information based on paths in the directed data lineage graph, and storing the summary information in one or more summary objects, the computing including

receiving designation of interest for a plurality of the nodes of the directed data lineage graph; and

generating one or more summary objects for remaining nodes not included in the plurality of nodes of interest, a first summary object of the one or more summary objects including summary information based on a first path between

a first node of interest and a second node of interest that does include one or more of the remaining nodes and does not include any nodes of interest other than the first and second nodes;

wherein generating the first summary object includes traversing the first path between the first node of interest and the second node of interest to identify one or more remaining nodes that are not designated as being of interest along the first path, and forming the summary information for the identified one or more remaining nodes.

20. The software of claim 19 wherein generating the first summary object includes traversing the first path between the first node of interest and the second node of interest to identify one or more remaining nodes that are not designated as being of interest along the first path, and forming the summary information for the identified one or more remaining nodes.

21. The software of claim 19 further including generating a data lineage diagram as a visual representation of at least portions of the directed data lineage graph, in which each node designated as being of interest is represented by its own icon, and a plurality of the remaining nodes are represented by a common icon connected to a plurality of nodes of interest according to summary information stored in one or more of the summary objects.

22. The software of claim 21 further including receiving a user input indicative of a user's desire to view details associated with the common icon and, in response to the user input, traversing the plurality of remaining nodes represented by the common icon and visually representing the remaining nodes in the data lineage diagram based on the traversal of paths of the directed data lineage graph including the plurality of remaining nodes associated with the summary information.

23. The software of claim 21 further including receiving a designation of one of the one or more data

transformation nodes or one of the one or more data nodes as a target node,

wherein generating the data lineage diagram includes traversing one or more paths through the directed data lineage graph, each path of the one or more paths including two or more nodes of interest and the target node.

24. The software of claim 23 wherein for at least some of the one or more paths through the directed data lineage graph, traversing the path includes traversing a first portion of the path in an

upstream direction from the target node and traversing a second portion of the path in a downstream direction from the target node.

25. The software of claim 24 wherein traversing the first portion of the path includes determining whether the target node is marked as being a node of interest, and

if the target node is determined to be a node of interest, traversing a link corresponding to a summary object associated with the target node from the target node to a first upstream node of interest, otherwise

if the target node is not determined to be a node of interest, traversing an original path of the directed data lineage graph between the target node and the first upstream node of interest, including traversing one or more data transformation nodes or data nodes that are not designated as being of interest.

26. The software of claim 25 wherein traversing the first portion of the path further includes traversing a link corresponding to a summary object associated with the first upstream node of interest from the first upstream node of interest to a second upstream node of interest.

27. The software of claim 24 wherein traversing the second portion of the path includes determining whether the target node is marked as being a node of interest, and

if the target node is determined to be a node of interest, traversing a link corresponding to a summary object associated with the target node from the target node to a first downstream node of interest, otherwise

if the target node is not determined to be a node of interest, traversing an original path of the directed data lineage graph between the target node and the first downstream node of interest, including traversing one or more data transformation nodes or data nodes that are not designated as being of interest.

28. The software of claim 27 wherein traversing the second portion of the path further includes traversing a link corresponding to a summary object associated with the first downstream node of interest from the first downstream node of interest to a second downstream node of interest.

29. The software of claim 21 wherein the one or more summary objects includes two or more summary objects, generating the lineage diagram further includes merging at least some of the

two or more summary objects into a summary node represented by the common icon.

30. The software of claim 29 wherein merging at least some of the two or more summary objects into the summary node includes analyzing data lineage relationships between the nodes of interest linked by the at least some of the two or more summary objects to determine whether merging the at least some of the two or more summary objects is possible.

31. The software of claim 30 wherein analyzing data lineage relationships between the nodes of interest includes determining that the at least some of the two or more summary objects are associated with a common downstream node of interest.

32. The software of claim 21 wherein the computing system includes a plurality of separate, interconnected sub-systems, the portions of the directed data lineage graph being distributed among at least some of the sub-systems.

33. The software of claim 32 wherein at least some of the sub-systems are represented by the common icon.

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

an input device or port for receiving an identification of a directed data lineage graph that includes one or more data transformation nodes that represent computations performed by the computing system that receive data as input, transform the received data, and produce the transformed data as output, and one or more data nodes that represent data including the data received as input and the data produced as output, and includes directed links that represent respective data lineage relationships between a computation and data to be received or produced by the computation during execution of the computation; and

at least one processor for computing summary information based on paths in the directed data lineage graph, and storing the summary information in one or more summary objects, the computing including

receiving designation of interest for a plurality of the nodes of the directed data lineage graph; and generating one or more summary objects for remaining nodes not included in the plurality of nodes of interest, a first summary object of the one or more summary

objects including summary information based on a first path between a first node of interest and a second node of interest that does include one or more of the remaining nodes and does not include any nodes of interest other than the first and second nodes;

wherein generating the first summary object includes traversing the first path between the first node of interest and the second node of interest to identify one or more remaining nodes that are not designated as being of interest along the first path, and forming the summary information for the identified one or more remaining nodes.

35. The computing system of claim 34 wherein generating the first summary object includes traversing the first path between the first node of interest and the second node of interest to identify one or more remaining nodes that are not designated as being of interest along the first path, and forming the summary information for the identified one or more remaining nodes.

36. The computing system of claim 34 further including generating a data lineage diagram as a visual representation of at least portions of the directed data lineage graph, in which each node designated as being of interest is represented by its own icon, and a plurality of the remaining nodes are represented by a common icon connected to a plurality of nodes of interest according to summary information stored in one or more of the summary objects.

37. The computing system of claim 36 further including receiving a user input indicative of a user's desire to view details associated with the common icon and, in response to the user input, traversing the plurality of remaining nodes represented by the common icon and visually representing the remaining nodes in the data lineage diagram based on the traversal of paths of the directed data lineage graph including the plurality of remaining nodes associated with the summary information.

38. The computing system of claim 36 further including receiving a designation of one of the one or more data

transformation nodes or one of the one or more data nodes as a target node,

wherein generating the data lineage diagram includes traversing one or more paths through the directed data lineage graph, each path of the one or more paths including two or more nodes of interest and the target node.

39. The computing system of claim 38 wherein for at least some of the one or more paths through the directed data lineage graph, traversing the path includes traversing a first portion of the

path in an upstream direction from the target node and traversing a second portion of the path in a downstream direction from the target node.

40. The computing system of claim 39 wherein traversing the first portion of the path includes determining whether the target node is marked as being a node of interest, and

if the target node is determined to be a node of interest, traversing a link corresponding to a summary object associated with the target node from the target node to a first upstream node of interest, otherwise

if the target node is not determined to be a node of interest, traversing an original path of the directed data lineage graph between the target node and the first upstream node of interest, including traversing one or more data transformation nodes or data nodes that are not designated as being of interest.

41. The computing system of claim 40 wherein traversing the first portion of the path further includes traversing a link corresponding to a summary object associated with the first upstream node of interest from the first upstream node of interest to a second upstream node of interest.

42. The computing system of claim 39 wherein traversing the second portion of the path includes determining whether the target node is marked as being a node of interest, and

if the target node is determined to be a node of interest, traversing a link corresponding to a summary object associated with the target node from the target node to a first downstream node of interest, otherwise

if the target node is not determined to be a node of interest, traversing an original path of the directed data lineage graph between the target node and the first downstream node of interest, including traversing one or more data transformation nodes or data nodes that are not designated as being of interest.

43. The computing system of claim 42 wherein traversing the second portion of the path further includes traversing a link corresponding to a summary object associated with the first downstream node of interest from the first downstream node of interest to a second downstream node of interest.

44. The computing system of claim 36 wherein the one or more summary objects includes two or more summary objects, generating the lineage diagram further includes merging at least some of the two or more summary objects into a summary node represented by the common icon.

45. The computing system of claim 44 wherein merging at least some of the two or more summary objects into the summary node includes analyzing data lineage relationships between the nodes of interest linked by the at least some of the two or more summary objects to determine whether merging the at least some of the two or more summary objects is possible.

46. The computing system of claim 45 wherein analyzing data lineage relationships between the nodes of interest includes determining that the at least some of the two or more summary objects are associated with a common downstream node of interest.

47. The computing system of claim 36 wherein the computing system includes a plurality of separate, interconnected sub-systems, the portions of the directed data lineage graph being distributed among at least some of the sub-systems.

48. The computing system of claim 47 wherein at least some of the sub-systems are represented by the common icon.

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

means for receiving an identification of a directed data lineage graph that includes one or more data transformation nodes that represent computations performed by the computing system that receive data as input, transform the received data, and produce the transformed data as output, and one or more data nodes that represent data including the data received as input and the data produced as output, and includes directed links that represent respective data lineage relationships between a computation and data to be received or produced by the computation during execution of the computation; and

means for computing summary information based on paths in the directed data lineage graph, and storing the summary information in one or more summary objects, the computing including

receiving designation of interest for a plurality of the nodes of the directed data lineage graph; and

generating one or more summary objects for remaining nodes not included in the

plurality of nodes of interest, a first summary object of the one or more summary objects including summary information based on a first path between a first node of interest and a second node of interest that does include one or more of the remaining nodes and does not include any nodes of interest other than the first and second nodes;

wherein generating the first summary object includes traversing the first path between the first node of interest and the second node of interest to identify one or more remaining nodes that are not designated as being of interest along the first path, and forming the summary information for the identified one or more remaining nodes.