CLAIMS

1. A method including:

at a node of a ~~Hadoop~~ cluster that stores a collection of data that can be operated on in parallel by nodes operating in conjunction with one another to carry out data processing operations on the collection of data, the node storing a first portion of data of the collection of data, ~~in HDFS data storage:~~

executing, at the node, a first instance of a data processing engine capable of accessing the first portion of data stored at the node and receiving data from a data source external to the ~~Hadoop~~ cluster;

receiving a ~~dataflow graph~~computer program by the data processing engine, the ~~dataflow graph~~computer program configured for accessing the first portion of data stored at the node and including a) at least one component representing the ~~Hadoop~~ cluster, b) at least one component representing the data source external to the ~~Hadoop~~ cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation;

executing at least part of the ~~dataflow graph~~computer program by the first instance of the data processing engine;

accessing, by the data processing engine, the first portion of data stored at the node;

receiving, by the data processing engine, a second portion of data from the external data source; and

performing, by the data processing engine, the data processing operation using at least the first portion of data and the second portion of data.

2. The method of claim 1 in which the ~~Hadoop cluster includes~~ nodes each ~~executing~~ execute an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel.

3. The method of claim 2 in which the data processing operation is performed on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by the other nodes of the ~~Hadoop~~ cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the ~~Hadoop~~ cluster.

~~3~~4. The method of claim 1 in which at least one component of the ~~dataflow graph~~computer program is connected to a link representing a flow of data from the ~~Hadoop~~ cluster, and wherein the at least one component is connected to a link representing a flow of data from the source of the second portion of data.

~~4~~5. The method of claim 1 in which at least some data of the collection of data is stored redundantly in multiple nodes~~the data processing engine does not implement the MapReduce programming model~~.

~~5~~6. The method of claim 1, in which the second portion of data is stored in volatile memory.

~~6~~7. The method of claim 1 including

receiving a database query, the database query including at least one operation to be performed on data received from at least one source of data that includes the ~~Hadoop~~ cluster; and

the ~~dataflow graph~~computer program includes components representing operations corresponding to the database query, wherein the ~~dataflow graph~~ computer program includes at least one component representing the at least one source of data and at least one link that represents at least one dataflow associated with the operation to be performed on data received from at least one source of data.

~~7~~8. The method of claim 1 in which the second portion of data was chosen based on characteristics of the first portion of data.

~~8~~9. The method of claim 1 in which the second portion of data includes a subset of rows of a relational database, and the second portion of data includes a subset of columns of the relational database.

~~9~~10. The method of claim 1 in which the second portion of data is distinct from a third portion of data received ~~at a second node of the Hadoop cluster~~ from the external data source at another node of the cluster.

~~10~~11. The method of claim 1 including communicating with an instance of at least part of the ~~dataflow graph~~ computer program that is being executed by a second instance of the data processing engine that is outside of the ~~Hadoop~~ cluster.

~~11~~12. The method of claim 1 including executing at least part of the dataflow graph by a second instance of the data processing engine outside of the ~~Hadoop~~ cluster.

13. The method of claim 1 in which the first portion of data is stored in distributed data storage.

14. The method of claim 1 in which the first portion of data is stored in HDFS data storage.

~~12~~15. A computer-readable storage device including instructions ~~for causing~~ which, when executed at a node of a ~~Hadoop~~ cluster storing a collection of data that can be operated on in parallel by nodes operating in conjunction with one another to carry out data processing operations on the collection of data, the node storing a first portion of data of the collection of data, cause the node to perform the method of any of claims 1 through 14.~~storing a first portion of data in HDFS data storage to carry out operations including:~~

~~executing a first instance of a data processing engine capable of receiving data from a data source external to the Hadoop cluster;~~

~~receiving a dataflow graph by the data processing engine, the dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing~~

~~the data source external to the Hadoop cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation;~~

~~executing at least part of the dataflow graph by the first instance of the data processing engine;~~

~~receiving, by the data processing engine, a second portion of data from the external data source;~~

~~and~~

~~performing, by the data processing engine, the data processing operation using at least the first portion of data and the second portion of data.~~

~~13~~16. A node of a ~~Hadoop~~ cluster that stores a collection of data that can be operated on in parallel by nodes operating in conjunction with one another to carry out data processing operations on the collection of data, the node storing a first portion of data of the collection of data, ~~in HDFS storage and~~ the node including a computer processing device configured to perform the method of any of claims 1 through 14.~~carry out operations including:~~

~~executing a first instance of a data processing engine capable of receiving data from a data source external to the Hadoop cluster;~~

~~receiving a dataflow graph by the data processing engine, the dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing the data source external to the Hadoop cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation;~~

~~executing at least part of the dataflow graph by the first instance of the data processing engine;~~

~~receiving, by the data processing engine, a second portion of data from the external data source;~~

~~and~~

~~performing, by the data processing engine, the data processing operation using at least the first portion of data and the second portion of data.~~

~~14~~17. A node of a ~~Hadoop~~ cluster that stores a collection of data that can be operated on in parallel by nodes operating in conjunction with one another to carry out data processing operations on the collection of data, the node storing a first portion of data of the collection of data, the node storing a first portion of data of the collection of data ~~in HDFS storage~~ and including:

means for executing a first instance of a data processing engine capable of accessing the first portion of data stored at the node and receiving data from a data source external to the ~~Hadoop~~ cluster;

means for receiving a ~~dataflow graph~~ computer program by the data processing engine, the ~~dataflow graph~~computer program configured for accessing the first portion of data stored at the node and including a) at least one component representing the ~~Hadoop~~ cluster, b) at least one component representing the data source external to the ~~Hadoop~~ cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation;

means for executing at least part of the ~~dataflow graph~~ computer program by the first instance of the data processing engine;

means for accessing, by the data processing engine, the first portion of data stored at the node;

means for receiving, by the data processing engine, a second portion of data from the external data source; and

means for performing, by the data processing engine, the data processing operation using at least the first portion of data and the second portion of data.

~~15~~18. A method including:

at a node of a cluster of nodes, the node storing a first portion of data and configured to carry out one or more data processing operations in conjunction with the cluster of nodes, the cluster storing a collection of data across the nodes, the nodes being configured to operate on the collection of data in parallel, the collection of data being split into portions, each portion being operated on by a respective node of the cluster:

executing a first instance of a data processing engine capable of receiving data from a data source external to the cluster;

receiving a dataflow graph by the data processing engine, the dataflow graph including a) at least one component representing the ~~Hadoop~~ cluster, b) at least one component representing the data source external to the ~~Hadoop~~ cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation;

executing at least part of the dataflow graph by the first instance of the data processing engine;

based on characteristics of the first portion of data, requesting, by the data processing engine, a second portion of data;

receiving, by the data processing engine, the second portion of data from the external data source;

storing the second portion of data in volatile memory of the node; and

performing, by the data processing engine, the data processing operation using at least the first portion of data and the second portion of data.

19. The method of claim 18 in which the cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by other nodes of the cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the cluster.

20. The method of claim 18 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the cluster, and wherein the at least one component is connected to a link representing a flow of data from the external data source of the second portion of data.

~~16~~21. A method including:

at a data processing engine of a node of a Hadoop cluster, performing a data processing operation identified by a dataflow graph being executed by the data processing engine, the

dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing the data source external to the Hadoop cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation, the data processing operation being performed using at least a first portion of data stored in HDFS data storage at the node and at least a second portion of data received from a data source external to the Hadoop cluster.

22. The method of claim 21 in which the Hadoop cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by other nodes of the Hadoop cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the Hadoop cluster.

23. The method of claim 21 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the Hadoop cluster, and wherein the at least one component is connected to a link representing a flow of data from the external data source of the second portion of data.

~~17~~24. A method including:

receiving a SQL query specifying sources of data including a Hadoop cluster and a relational database;

generating a dataflow graph that corresponds to the SQL query, the dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing the data source external to the Hadoop cluster, and c) at least one link that represents at least one dataflow associated with a data processing operation;

executing the dataflow graph at a data processing engine of a node of the Hadoop cluster; and

performing, by the data processing engine, the data processing operation using at least data of the Hadoop cluster and data of the relational database.

~~18. The non-transitory computer-readable storage device of claim 12 in which the Hadoop cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by other nodes of the Hadoop cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the Hadoop cluster.~~

~~19. The non-transitory computer-readable storage device of claim 12 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the Hadoop cluster, and wherein the at least one component is connected to a link representing a flow of data from the external data source of the second portion of data.~~

~~20. The non-transitory computer-readable storage device of claim 12 in which the data processing engine does not implement a MapReduce programming model.~~

~~21. The non-transitory computer-readable storage device of claim 12, in which the second portion of data is stored in volatile memory.~~

~~22. The non-transitory computer-readable storage device of claim 12, the operations including~~

~~receiving a database query, the database query including at least one operation to be performed on data received from at least one source of data that includes the Hadoop cluster; and~~

~~the dataflow graph including components representing operations corresponding to the database query, wherein the dataflow graph includes at least one component representing the at~~

~~least one source of data and at least one link that represents at least one dataflow associated with the operation to be performed on data received from at least one source of data.~~

~~23. The non-transitory computer-readable storage device of claim 12 in which the second portion of data was chosen based on characteristics of the first portion of data.~~

~~24. The non-transitory computer-readable storage device of claim 12 in which the second portion of data includes a subset of rows of a relational database, and the second portion of data includes a subset of columns of the relational database.~~

~~25. The non-transitory computer-readable storage device of claim 12 in which the second portion of data is distinct from a third portion of data received at a second node of the Hadoop cluster from the external data source.~~

~~26. The non-transitory computer-readable storage device of claim 12, the operations including communicating with an instance of at least part of the dataflow graph that is being executed by a second instance of the data processing engine that is outside of the Hadoop cluster.~~

~~27. The non-transitory computer-readable storage device of claim 12, the operations including executing at least part of the dataflow graph by a second instance of the data processing engine outside of the Hadoop cluster.~~

~~28. The node of claim 13 in which the Hadoop cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by other nodes of the Hadoop cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the Hadoop cluster.~~

29. The node of claim 13 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the Hadoop cluster, and wherein the at least one component is connected to a link representing a flow of data from the external data source of the second portion of data.

30. The node of claim 13 in which the data processing engine does not implement a MapReduce programming model.

31. The node of claim 13, in which the second portion of data is stored in volatile memory.

32. The node of claim 13, the operations including

receiving a database query, the database query including at least one operation to be performed on data received from at least one source of data that includes the Hadoop cluster; and

the dataflow graph including components representing operations corresponding to the database query, wherein the dataflow graph includes at least one component representing the at least one source of data and at least one link that represents at least one dataflow associated with the operation to be performed on data received from at least one source of data.

33. The node of claim 13 in which the second portion of data was chosen based on characteristics of the first portion of data.

34. The node of claim 13 in which the second portion of data includes a subset of rows of a relational database, and the second portion of data includes a subset of columns of the relational database.

35. The node of claim 13 in which the second portion of data is distinct from a third portion of data received at a second node of the Hadoop cluster from the external data source.

~~36. The node of claim 13, the operations including communicating with an instance of at least part of the dataflow graph that is being executed by a second instance of the data processing engine that is outside of the Hadoop cluster.~~

~~37. The node of claim 13, the operations including executing at least part of the dataflow graph by a second instance of the data processing engine outside of the Hadoop cluster.~~

~~38. The method of claim 15 in which the Hadoop cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by other nodes of the Hadoop cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the Hadoop cluster.~~

~~39. The method of claim 15 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the Hadoop cluster, and wherein the at least one component is connected to a link representing a flow of data from the external data source of the second portion of data.40. The method of claim 16 in which the Hadoop cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes the first portion of data, the first body of data also including other portions of data being processed by other nodes of the Hadoop cluster, and b) a second body of data that includes the second portion of data, the second body of data being stored in a format native to a relational database system, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the Hadoop cluster.~~

~~41. The method of claim 16 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the Hadoop cluster, and wherein the at least~~

~~one component is connected to a link representing a flow of data from the external data source of the second portion of data.~~

~~42~~25. The method of claim ~~17~~ 24 in which the Hadoop cluster includes nodes each executing an instance of the data processing engine, the instances of the data processing engine running concurrently to perform the data processing operation together in parallel on a) a first body of data that includes a first portion of data, the first body of data also including other portions of data being processed by other nodes of the Hadoop cluster, and b) a second body of data that includes a second portion of data, the second body of data being stored in a format native to the relational database, and the second body of data being divided into portions that each can be stored in volatile memory of the nodes of the Hadoop cluster.

~~43~~26. The method of claim ~~17~~ 24 in which at least one component of the dataflow graph is connected to a link representing a flow of data from the Hadoop cluster, and wherein the at least one component is connected to a link representing a flow of data from the relational database.