

What is claimed is:

1. A method including;

at a node of a Hadoop cluster, the node storing a first portion of data in HDFS data storage:

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

receiving a computer-executable program by the data processing engine, the computer-executable program including a dataflow graph capable of being executed by a graph execution engine of the data processing engine;

executing at least part of the computer-executable program 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;

storing the second portion of data other than in HDFS storage; and

performing, by the data processing engine, a data processing operation identified by the computer-executable program using at least the first portion of data and the second portion of data,

the dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing the external data source, and c) at least one link that represents at least one dataflow associated with the operation to be performed on the 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 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.

3. The method of claim 1 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.

4. The method of claim 1 in which the data processing engine does not implement a MapReduce programming model.

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

6. 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 computer program includes components representing operations corresponding to the database query, wherein the 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. The method of claim 1 in which the second portion of data was chosen based on characteristics of the first portion of data.

8. 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. 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.

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

11. The method of claim 1 including executing at least part of the computer-executable program by a second instance of the data processing engine outside of the Hadoop cluster.

12. A non-transitory computer-readable storage device including instructions for causing a node of a Hadoop cluster storing a first portion of data in HDFS data storage to carry out the steps of the method of any one of claims 1 to 11.

13. A node of a Hadoop cluster storing a first portion of data in HDFS data storage and including a computer processing device configured to carry out the steps of the method of any one of claims 1 to 11.

14. A node of a Hadoop cluster storing a first portion of data in HDFS storage and including a computer processing device and the node including: means for executing a first instance of a data processing engine capable of receiving data from a data source external to the Hadoop cluster; means for receiving a program by the data processing engine, the program including a dataflow graph capable of being executed by a graph execution engine of the data processing engine; means for executing at least part of the program by the first instance of the data processing engine; means for receiving, by the data processing engine, a second portion of data from the external data source; means for storing the second portion of data other than in HDFS storage; and means for performing, by the data processing engine, a data processing operation identified by the program using at least the first portion of data and the second portion of data, the dataflow graph

including a) at least one component representing the Hadoop cluster, b) at least one component representing the external data source, and c) at least one link that represents at least one dataflow associated with the operation to be performed on the at least the first portion of data and the second portion of data.

15. A method including: at a node storing a first portion of data and operating in conjunction with a Hadoop cluster of nodes, the cluster storing an aggregation of data, the nodes being configured to operate on the aggregation of data in parallel: executing a first instance of a data processing engine capable of receiving data from a data source external to the Hadoop cluster; receiving a computer-executable program by the data processing engine, the computer-executable program including a dataflow graph capable of being executed by a graph execution engine of the data processing engine; executing at least part of the computer-executable program 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; storing the second portion of data in volatile memory of the node; and performing, by the data processing engine, a data processing operation identified by the computer-executable program using at least the first portion of data and the second portion of data, the dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing the external data source, and c) at least one link that represents at least one dataflow associated with the operation to be performed on the at least the first portion of data and the second portion of data.

16. 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.

17. The method of claim 15 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.

18. A method including: at a data processing engine of a node of a Hadoop cluster, performing a data processing operation identified by a computer-executable program being executed by the data processing engine, the computer-executable program including, a dataflow graph capable of being executed by a graph execution engine of the data processing engine, 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 and stored other than in HDFS storage, the dataflow graph including a) at least one component representing the Hadoop cluster, h) 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 the operation performed on the at least the first portion of data and the second portion of data.

19. The method of claim 18 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.

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 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.

21. A method including: receiving a SQL query specifying sources of data including a Hadoop cluster and a relational database; generating a computer-executable program that corresponds to the SQL query, the computer-executable program including a dataflow graph capable of being executed by a graph execution engine of the data processing engine; executing the computer-executable program at a data processing engine of a node of the Hadoop cluster; and performing, by the data processing engine, a data processing operation identified by the computer-executable program using at least data of the Hadoop cluster and data of the relational database, the dataflow graph including a) at least one component representing the Hadoop cluster, b) at least one component representing the relational database, and c) at least one link that represents at least one dataflow associated with the operation that uses the data of the Hadoop cluster and the data of the relational database.

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 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.

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 relational database.