

ABSTRACT

PROCESSING DATA FROM MULTIPLE SOURCES

In a first aspect, a method includes, 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, executing at least part of the 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 program using at least the first portion of data and the second portion of data.

(Fig. 1)