

Quiz #2: MapReduce Week 2 Name: _____ ID: _____

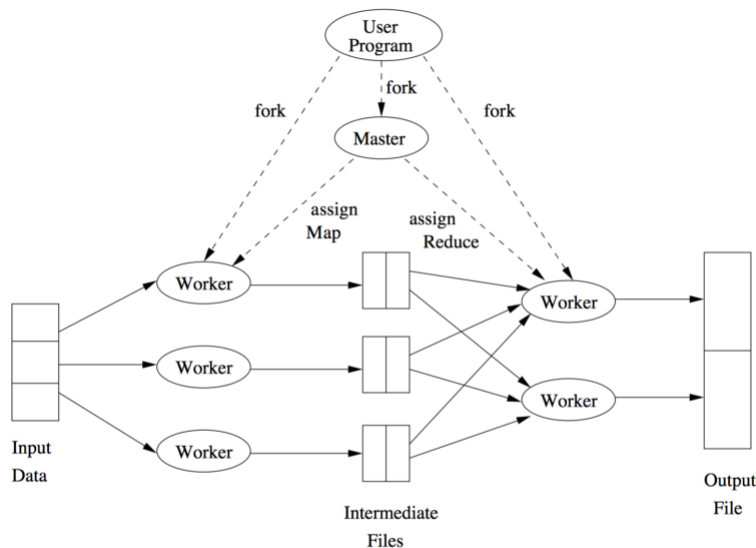


Figure 2.3: Overview of the execution of a MapReduce program

Q1. Concerning the above figure, where does the MapReduce program store 1. Input data, 2. Intermediate files, and 3. Output data? (3 pts)

1. DFS/HDFS (1 pt)
2. Local FS (1 pt)
3. DFS/HDFS (1 pt)

Q2. Why is that when map workers fail, the tasks that are completed or in-progress at map workers are reset to idle and rescheduled but only in-progress tasks are reset to idle when reduce worker fails? (3pts)

When a map task completes, it sends the master the location and sizes of its intermediate files. If map worker fails, the intermediate result is no longer available to reducers. So both completed and in-progress task should be reset and rescheduled. (1.5 pts)

But reducer directly writes the results to file system. So only the in-progress task should be reset to idle when reducer fails. (1.5 pts)

Q3. Write the Map and Reduce tasks and their output for joining these two tables: (4pts)

Order(orderid, account, date)

1, aaa, d1
2, aaa, d2
3, bbb, d3


LineItem(orderid, itemid, qty)

1, 10, 1
1, 20, 3
2, 10, 5
2, 50, 100
3, 20, 1

(pseudo code or plain words are both fine.)

Map task: (2 pts)

Use orderid as key, and table name together with other columns as value. Map each row in the table and emit the key-value pair.

Map Task **relation name** 

Order

1, aaa, d1	→	1 : "Order", (1,aaa,d1)
2, aaa, d2	→	2 : "Order", (2,aaa,d2)
3, bbb, d3	→	3 : "Order", (3,bbb,d3)

Line

1, 10, 1	→	1 : "Line", (1, 10, 1)
1, 20, 3	→	1 : "Line", (1, 20, 3)
2, 10, 5	→	2 : "Line", (2, 10, 5)
2, 50, 100	→	2 : "Line", (2, 50, 100)
3, 20, 1	→	3 : "Line", (3, 20, 1)

Reduce task: (2 pts)

groups together all values (tuples) associated with each key and emit joined values.

Reducer for key 1

"Order", (1,aaa,d1)
"Line", (1, 10, 1)
"Line", (1, 20, 3)



(1, aaa, d1, 1, 10, 1)
(1, aaa, d1, 1, 20, 3)