WEEK 2 QUIZ SOLUTION

1. As a good developer where would you like to do more processing
   a. ***Mapper**
   b. Reducer
   c. Partitioner
   d. Record reader

   Explanation: Mapper because it's a first processing node. If maximum work is done in mapper that means less mapper output is generated, means less data shuffle and network bandwidth needed which will lead to good performance.

2. Within each mapper, rows are processed _____.
   a. Parallelly
   b. ***Sequentially**

   Explanation: Input file is logically divided into input split and every input split is assigned one mapper. Multiple mappers run in parallel. But within each mapper the lines inside input split are read sequentially.

3. Consider a scenario, where we have an 880 MB of input file and 3 node cluster, and we want to do some processing not involving aggregation. Which of these will be best suited?
   a. 6 Mappers, 1 Reducer
   b. 6 Mappers, 0 Reducer
   c. 7 Mappers, 1 Reducer
   d. ***7 Mappers, 0 Reducer**

   Explanation: 880/128=6.875=7 Mapper. There is no aggregation involved hence 0 reducers.

4. Which of these will determine which output key from mapper goes to which reducer, when number of reducers are more than one?
   a. Combiner
   b. ***Partitioner**

   Explanation: Partitioner uses hash function to decide which output key from mapper goes to which reducer

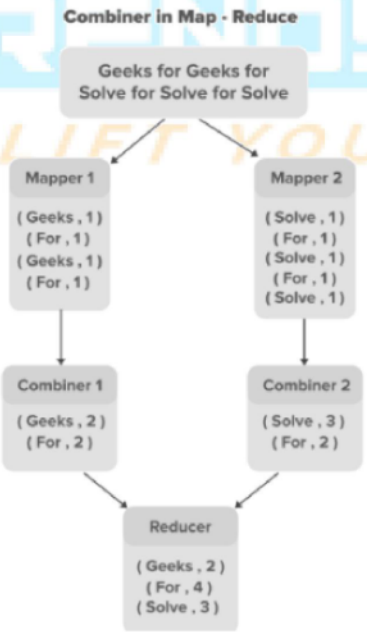
WEEK 2 QUIZ SOLUTION

5. Which of the following about hash functions is true? (Choose two)
   a. ***Same key goes to same reducer every time it occurs**
   b. Hash function code is fixed by framework and can't be changed
   c. Same key occurring multiple times will be assigned to a different reducer each time
   d. ***Developer can overwrite the default hash function code**

   Explanation: Hadoop provides default partitioner which has hash function, it can be overridden based on user requirement. It ensures same key goes to same reducer every time.

6. Which of the following holds false for Combiners? (choose two)
   a. Combiner logic may or may not be same as Reducer logic
   b. Combiners decrease the data transfer time and mappers work more
   c. ***Combiners increase the data transfer time and mappers work less**
   d. ***Combines values with different keys together**

   Explanation: C & D. Combiner does the job of reducer (aggregation of same key) for every mapper and hence the output generated by mapper is less, which decrease data transfer time.



WEEK 2 QUIZ SOLUTION

7. In which of these scenarios a Reducer does not serve any purpose?
   a. Google inverted index problem
   b. Aggregation problem
   c. Finding distinct search term from all searches problem
   d. *Filtering of rows

   Explanation: Filtering of rows just needs a condition in mapper. Mapper and reducer output will be same hence reducer is not required.
   A, B, & C needs aggregation and hence reducer is required.

8. In which of these cases the Reducer logic cannot be used as Combiner logic?
   a. Maximum Entity from a list of entities
   b. Finding Sum of entities
   c. *Finding Average of a list of entities
   d. Minimum Entity from a list of entities

   Explanation: Below image explains that combiner and reducer cannot have same logic for calculating average.

					Max	Min	Sum	Avg
Mapper 1	10	20	30	Combiner 1	30	10	60	20
Mapper 2	11	65	31	Combiner 2	65	11	107	35.66667
Mapper 3	12			Combiner 3	12	12	12	12
Mapper 4	13	23		Combiner 4	23	13	36	18
				Reducer	65	10	215	21.41667

					Max	Min	Sum	Avg
Mapper 1	10	20	30					
Mapper 2	11	65	31	Reducer	65	10	215	23.88889
Mapper 3	12							
Mapper 4	13	23						

9. Shuffle phase generally involves significant overhead of data transfer across network where reducers are involved, how can we optimize the MR program in such cases?
   a. Introduce more and more mappers to get greater parallelism
   b. Make number of reducers as zero
   c. *Introduce Combiners

   Explanation: To reduce shuffle we can do local aggregation using combiner instead of passing lot of key-value data.

WEEK 2 QUIZ SOLUTION

10. if you have a 500 mb file and you want 8 mappers to run on it, then what will you do
    a. Define the number of mappers in java Main class
    b. Define the number of mappers in java Map class
    c. *Change the block size to 64 mb when ingesting the file to hdfs
    d. None of the above

    Explanation: No of mappers is decided on file size / block size.
    E.g., 500/128 =4 and 500/64=8

11. Do all 3 replicas of a block execute in parallel?
    a. Yes
    b. *No

    Explanation: Only one nearest block is picked for execution.
    Replication of 3 helps for fault tolerance.

12. When running MapReduce program. If you provide the output directory path which already exists then what will happen?
    a. It will overwrite the output directory
    b. *The job will fail
    c. We can't say
    d. None of these

    Explanation: If the output directory exists at given path, then it will give error and job will be failed.

13. What if our hash function is not consistent?
    a. Job will error out
    b. *We might get same key in multiple reducers so final aggregation won't happen
    c. We will get correct output
    d. None of the above

    Explanation: If hash function is not consistent then same key will go to multiple reducers hence the same will be given in reducer output. In other words, Final aggregation is not done.

WEEK 2 QUIZ SOLUTION

14. Consider you have a 500 MB file. How to change number of reducers to 4
    a. The number of reducers will be 4 in this case as its a 500 mb file
    b. Change the block size to get 4 reducers
    c. *Specify the number of reducers in java Main class
    d. None of the above

    Explanation: No of reducer can be configured by developer in main program.

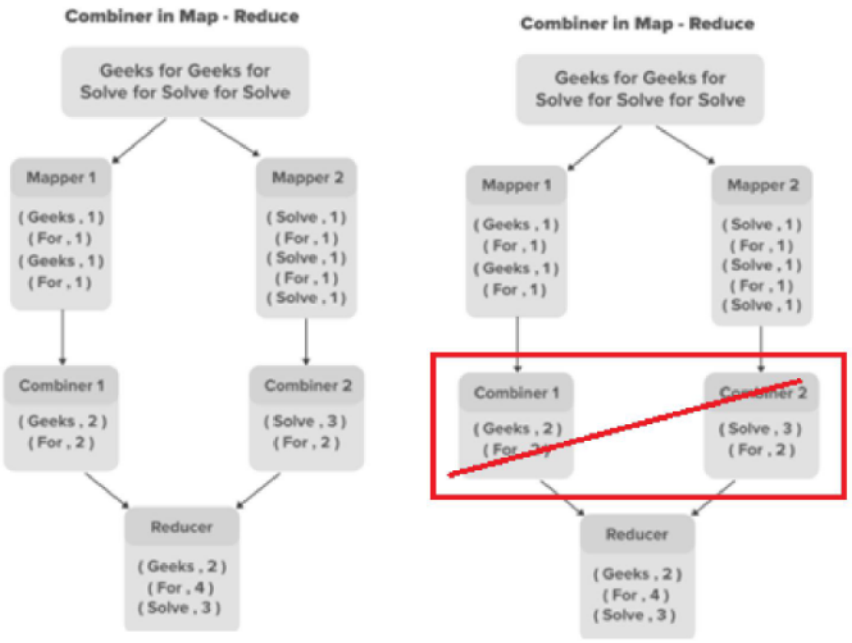
15. On which machine does combiner run
    a. Reducer machine
    b. *Mapper machine
    c. Both mapper and reducer machines
    d. none of the machine

    Explanation: combiner executes on mapper machine and hence it reduces the data transfer between mapper and reducer.

16. What is the role of combiner (select 2)?
    a. It reduces the load of mapper
    b. *It reduces the load of reducer
    c. *Less shuffling
    d. More shuffling

    Explanation: As you can see in the below diagram, with combiner only 4 key-value pair are transferred (less load on reducer as well as less shuffling).
    Without combiner only 9 key-value pair are transferred (more load on reducer as well as more shuffling).

WEEK 2 QUIZ SOLUTION



17. Is this a consistent function: if key.length + 2 < 5 then return 0 else return 1?
    a. ***Yes**
    b. No
    c. Can't say
    d. In some cases, yes and in other cases no

    Explanation: For same length word you will always get same result.
    For e.g.
    Key-one → 3+2<5 =1
    Key-Two → 3+2<5 =1
    Key-Four → 4+2<5 =0
    Key-Five → 4+2<5 =0

18. Can we process a directory with multiple files using MapReduce?
    a. ***Yes**
    b. No

    Explanation: Yes, when you pass directory path for input, it will process all files in that directory.

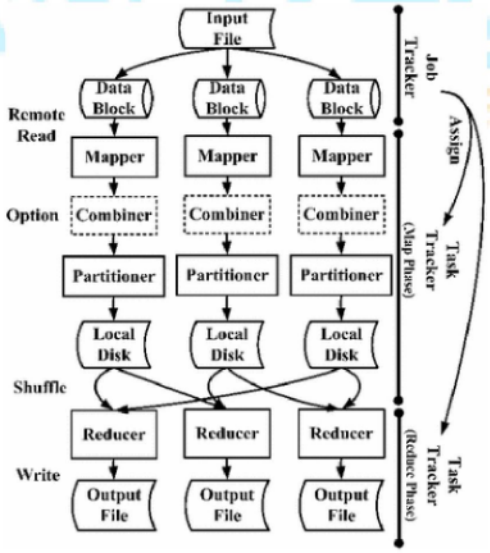
19. Consider you have a 500 MB file in HDFS. Number of reducers is set to 2. then how many part files will be created in output folder?

a. 1
b. 4
c. *2
d. none of the above
Explanation: No of reducer is equal to no of output files

20. Consider you have a 500 MB file in HDFS. Number of reducers is set to 0. then how many part files will be created in output folder.
    a. 0
    b. 1
    c. 2
    d. *4
    Explanation: Here there is no reducer so the output from mapper is final. No of mapper = file size/block size =500/128=4

21. How many combiners will work in MR Program?
    a. Equal to number of reducers
    b. Equal to number of partitioners
    c. We can set how many we want
    d. *Equal to number of mappers
    Explanation: Combiner works on mapper node hence no of mappers are equal to no of combiners.

e.

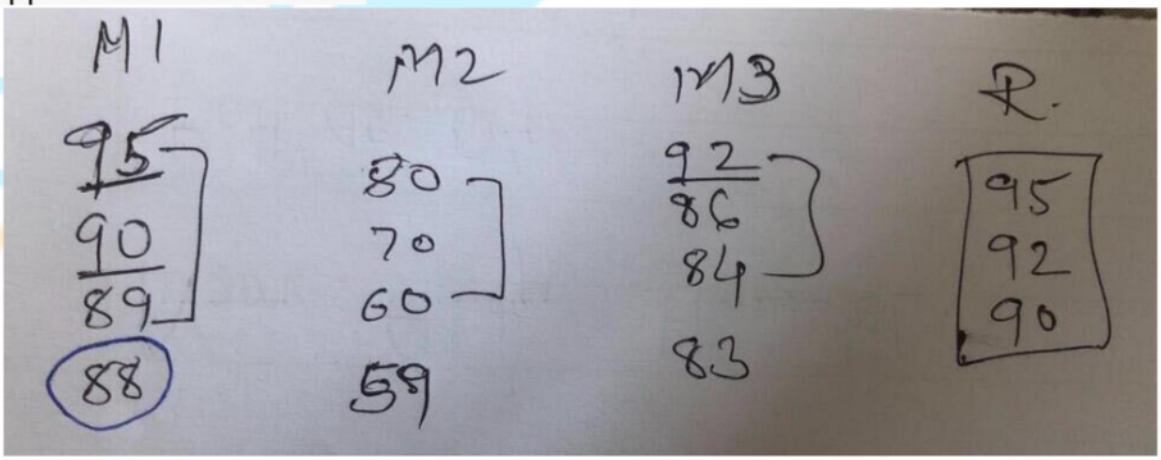


WEEK 2 QUIZ SOLUTION

22. We have 20 mappers, consider each mapper holds 1 lakh rows. if we want to find top 10 ranks out of these, then how many rows each mapper should emit?
    a. can be anything it depends on data
    b. 20 rows
    c. 1 row
    d. *10 rows

Explanation: Every mapper output top 10 records.
For 20 different mappers, each has to give their top 10 ranks. Then if you reduce those numbers (20* 10 = 200) then you'll get 200 values and now if you take the top 10, you'll get the top 10 rank among the entire data set.
In any case, no matter how many rows each mapper process, it can't have more than 10 top ranks. So even if a single mapper has all the top 10 records among the entire data set, you'll get the top 10 from the entire dataset in that way.

Below is the diagram explanation for finding top 3 marks using 3 mappers and 4 rows each.



WEEK 2 QUIZ SOLUTION

23. Choose the correct ones (multiple can be correct)
    a. ***Hash function should be consistent**
    b. ***We should make sure our hash function should evenly distribute the keys.**
    c. Hash function should not be consistent
    d. None of the above

    Explanation: While creating has function we must make sure that they are consistent and distributes data evenly.

    For e.g., 1: Mobile number starting with 0,1, 2 or 3 goes to partition1 and 4,5,6,7,8,9 goes to partition 2.
    This is consistent but not evenly distributed.
    For e.g., 2: Mobile number starting with 0, 1, 2, 3, 4 goes to partition1 and 5,6,7,8,9 goes to partition 2.
    This is consistent as well as evenly distributed.

    Note: even distribution is considered in natural terms at development time. At runtime in some case, it might not be evenly distributed.

24. Is there a possibility that we can have 1 mapper and 2 reducers?
    a. ***yes**
    b. no

    Explanation: Yes, I can do it practically but it does not make any sense in real scenarios. In real scenarios. Reducers are way less compared to mappers as its job is to do aggregation.

25. Where does partitioning happens?
    a. ***mapper machine**
    b. reducer machine

    Explanation: Partitioning happens on mapper machine. Refer diagram given for Q no 21 explanation.

WEEK 2 QUIZ SOLUTION

26. If you are filtering for something then how should your filename look like.
    a. part-r-00000
    b. *part-m-00000
    c. can be any of the above
    d. none of the above

    Explanation: As we saw in Q No 7 that filtering is a map only job and hence the output file name will be part-m-00000

27. Will shuffle & sort always happen?
    a. Yes, in every case it will happen
    b. *It won't happen when number of reducer is set 0
    c. We can't say
    d. None of the above

    Explanation: Shuffle and sort takes place only if we have reducer involved

28. What is a counter in MapReduce?
    a. It counts the number of words in mapper
    b. *It gives the statistics
    c. It counts the number of words in reducer
    d. None of the above

    Explanation: A Counter in MapReduce is a mechanism used for collecting and measuring statistical information about MapReduce jobs and events.

29. How do 2 reducers communicate with each other?
    a. via ssh
    b. using internal mechanism
    c. *they don't have to communicate
    d. none of the above

    Explanation: Every reducer output one file hence they don't need to communicate.

WEEK 2 QUIZ SOLUTION

30. Let's say there is a function which takes number of Dollars as input and convert it to Rs. Is this a consistent function.
    a. Yes
    b. *No
    c. We can't say

    Explanation: No, it's not consistent because dollar conversion rate keeps on changing every moment, hence input and output won't match always.

31. Which command is used to copy a file from location1 in HDFS to location2 in HDFS?
    a. hadoop fs -copyFromLocal location1 location2
    b. hadoop fs -put location1 location2
    c. *hadoop fs -cp location1 location2
    d. none of the above

    Explanation: -cp command is used to copy the file between HDFS