







Problem: Product popularity For e-commerce portalto suggest offers your client, Marketplace Inc., is a well established online retail company in US for consumer goods. Many users visit client’s website daily. The client wants to analyse the most popular products for each region and their popularity level sl that they can plan different offers to increase their sales. The input data is in single database table with collowing columns:

ProductId

UserId

User Location

Visit Time

We have a region table with following information

User Location

Region

We need to know the popularity of product in following format B:

Product Id

Region

Popularity

Each user is assigned to one Region. Size of region table is around 200 mb.

What you need to do:

Find popularity of each product for each region so that sales team can draft a discount strategy

Part 1 45 mins:

Please list out all the assumptions that you make clearly.

You are required to do the following

1. Load the data from Oracle/MySQL database to hadoop
2. Write MapReduce job to achieve the task
3. Please mention your design consideration while choosing Hive or HBase or HDFS for the task

Please list out all the assumption that you make clearly

Part2: 30 mins hadoop

Answer the following questions, each with one or more correct answers. For each checked option, answer in one line why you have chosen that option

1. In which of the following situation would you recommend an HBase tool to be used:
2. When updates are required for data
3. When we have variable schema problem
4. When we want randomly access data
5. B and C are correct
6. A,B, C are correct
7. None of the above
8. Where is FSImage and edit logs stored
9. Job tracker
10. Name node
11. Data node
12. Task tracker
13. Secondary name node
14. Which of the following is used for sampling in Hive
15. Partitions
16. Buckers
17. Filter
18. Splits
19. Samples
20. Using combiners has following advantages
21. Reduce execution time
22. We don’t have to run reducer
23. Reducer network traffic
24. Increase parallelism
25. \_\_\_\_\_\_are properties must for combiner to implement
26. In what case you will choose hive over java map reduce and vice versa
27. What is UDF and How will you create UDF in Pig
28. How will you overcome name node failover in Hadoop?
29. Why we are we able to access data faster from HDFS