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:

a.       When updates are required for data

b.      When we have variable schema problem

c.       When we want randomly access data

d.      B and C are correct

e.      A,B, C are correct

f.        None of the above

2.       Where is FSImage and edit logs stored

a.       Job tracker

b.      Name node

c.       Data node

d.      Task tracker

e.      Secondary name node

3.       Which of the following is used for sampling in Hive

a.       Partitions

b.      Buckers

c.       Filter

d.      Splits

e.      Samples

4.       Using combiners has following advantages

a.       Reduce execution time

b.      We don’t have to run reducer

c.       Reducer network traffic

d.      Increase parallelism

5.       \_\_\_\_\_\_are properties must for combiner to implement

6.       In what case you will choose hive over java map reduce and vice versa

7.       What is UDF and How will you create UDF in Pig

8.       How will you overcome name node failover in Hadoop?

9.       Why we are we able to access data faster from HDFS

https://ssl.gstatic.com/ui/v1/icons/mail/images/cleardot.gif

Updates will be not often for us to consider

2.       Where is FSImage and edit logs stored

a.       Job tracker

b.      Name node

c.       Data node

d.      Task tracker

e.      Secondary name node

Part of meta data

3.       Which of the following is used for sampling in Hive

a.       Partitions

b.      Buckers

c.       Filter

d.      Splits

e.      Samples

We can sample data based on bucketing table

4.       Using combiners has following advantages

a.       Reduce execution time

b.      We don’t have to run reducer

c.       Reducer network traffic

d.      Increase parallelism

Primary goal of combiner is to reduce the key/value pair shuffle across network

5.       \_\_\_\_\_\_are properties must for combiner to implement

Job.setCombinerClass(Combiner.class)

6.       In what case you will choose hive over java map reduce and vice versa

Hive for structured data and map reduce for unstructured data

7.       What is UDF and How will you create UDF in Pig

UDF – User Defined Function. We need to write Java function program which contains logic. And make it jar file and register in the pig program by defining in the initial lines.

8.       How will you overcome name node failover in Hadoop?

By adding secondary node

9.       Why we are we able to access data faster from HDFS

Hdfs runs on cluster setup which holds data on many nodes with replication factor. By running parallel hdfs can be able to provide the data in redundant manner.