1.       Explain About HDFS Blocks

2.       HDFS Blocks and RAM memory consideration and different scenarios on block size and memory size

3.       Explain Job Tracker

4.       How you worked with Structured Data and Unstructured data. More explanation on Unstructured data how you consider writing Map-Reduce program from log files.

5.       How you handled unstructured data in Hive?

6.       How you validate unstructured data (data type) in Map-Reduce

7.       Explain about Partitions in Hadoop

8.       Explain partitions in Hive and you use them

9.       Explain debugging process in Hive

10.   Different Joins in Hive

11.   Difference between right join and outer join in Hive

12.   Difference between left semi join and inner join

13.   How you done serialization in Hadoop

Can you please walk me through the following scenario how it is done in

Hadoop Map Reduce and what happens internally?  
  
Input File: Employee\_Sal.csv (File attached)  
O/P:  
Sum(Salary)  
Average(Salary)  
Count(ID)  
Avg\_Sal\_Month = Average(Salary/12)  
  
First row in the file has column names.

2.How to avoid shuffling in the mapping o/p? When I try to print integers 1 to 10 in ascending order, map o/p is printing in some weird order, why?

3.   In the mapping output of reverse string, o/p is pair. How to get just the ‘key’ value?

a.       Example:

Input File:

                                    rama.txt

Output File:

                                  revstrop/par\*

niassuh rim     0

-- How to avoid this 0 in the part-0000 file??

-- We have not used Reducer class here?

3. Where is this map reduce programming used in real time and when sqoop/hive/pig are used? Which one to choose when?

4.Like what was the source? How many files? How did you load into HDFS? Design and development, what requirements? How did you deliver? Etc…

1.      Hadoop API –

**Is it same what we are using? Mapper Class, Reducer Class and Configuration class**

2.      What is Streaming API? Have you used any of that?

3.      What is Driver Code?

**– Is it same as Configuration class?**

4.      Input file Format types – PDF? Zipped File? not splittable file format?

5.      When to use Partitioner? And how to determine the optimal number of Reducers for a job?

6.      what is a secondary sort? How to snaffle work in mapreduce?

7.      How to Indexing data?

8.      Can you join datasets using Map Reduce jobs? Map Side Join? Reduce-Side join?

9.      Is there any intermediate file generated after mapper (internal)? Where is it created local or hdfs? How can we see it?

10.  What debug tool did you use?

SCENARIO:

iam developing a Hadoop app

I have a lot of files composed by data about daily global weather (**like the \*.op attached file**) and I need to extract some information of these files.

I've achieved the maximum temperature of a year based on the Hadoop Definitive Guide, 2nd edition.

Now I have to sum the occurences of all phenomena in each year by station (the last 6 fields of each line indicates the occurence or not of the phenomena).

The description of the data format is in the GSOD\_DESC.txt.

CAN  U SOLVE ...

 I loaded the data into a hive table and I am able to query max temp, max dewp etc. I had to convert the data file to csv and remove the header information.

Here is the table I created

CREATE TABLE weatherdata (STN INT, WBAN INT, YEARMODA INT, TEMP FLOAT, TEMPDUR INT, DEWP FLOAT, DEWPDUR INT, SLP FLOAT, SLPDUR INT, STP FLOAT, STPDUR INT, VSIB FLOAT, VSIBDUR INT, WDSP FLOAT, WDSPDUR INT, MXSPD FLOAT, GUST FLOAT, MAX FLOAT, MIN FLOAT, PRCP FLOAT, SNDP FLOAT, FRSHTT INT)row format delimited fields terminated by ',' escaped by '\\';

This is how I loaded the CSV file into the table

load data inpath 'WEATHER/WeatherData.csv' into table weatherdata;

This are some queries I am able to run

select max(temp) from weatherdata;

select max(dewp) from weatherdata;

Let me know what else I need to do. Thanks a lot for this project. I am feeling very comfortable with hive now.