Amex Interview questions:

1. Write a MR program to calculate mean for one TB of data with integers?  
   i) Write logic to parse the data which separates integers in Mapper class  
   ii) Have a static/customer key for all the integers, so that values will be passed to a single reducer to calculate mean.  
   iii) write logic to calculate mean inside the reducer
2. Write a MR program to calculate Average for one TB of data with integers?  
   i) Write logic to parse the data which separates integers in Mapper class  
    use string tokenizer to separate the values  
   ii) take 2 global variables **count** and **sum** which is initiated to zero.  
   iii) Iterate the values with in the for loop and increment **count** inside the loop and calculate the sum.  
   iv) Now we have sum and count values, so we can calculate average as (sum/count)
3. How will you debug, if average is not performing?  
   i) Introduce logs/sysouts, where ever is required.  
   ii) use counters to cross check logs
4. How will you debug and identify the issue, if average is not performing without changing the code?  
   i) Identify the job id and check for logs and java exceptions.  
   ii) Mappers output will be stored with in the local system, instead of passing one TB of data, pass small amount of data(1mb) and test whether average is performing well or not. If average is performing well for I mb of data, same case with one TB.
5. How will you write a map reduce program to calculate Average without using reducer?  
   Implement mapper and reducer steps with in the mapper class.
6. How to avoid duplicate values in mapper/reducer?  
    out put from the mapper may have duplicates too, pass this list to set collection type to avoid duplicates in the reducer class.
7. Difference between Hive and mongodb?

Hive:

Hive is similar to SQL, but all SQL commands won’t work in Hive   
We can only read already existing files in HDFS in Hive, It won’t support UPDATE or DELETE.   
All hive queries will convert into MR’s and perform the result.

Mongodb:

Mongodb is noSQL, which can perform all the operations as database.  
Mongodb doesn’t have schema  
we can insert n number of columns  
MongoDB, collections, documents and fields are maintained in the database

1. Difference between Hive and pig?  
   Hive:

Hive is similar to SQL, but all SQL commands won’t work in Hive   
We can only read already existing files in HDFS in Hive, It won’t support UPDATE or DELETE.   
All hive queries will convert into MR’s and perform the result.

Pig:  
Pig is a scripting language and maintain datatypes and schemas within the scripts.   
Pig has data transformation functions  
Pig has datatypes include sets, associative arrays, tuples  
Pig is developed at yahoo!

Hive and Pig have support for joining, ordering or sorting data dynamically

1. Have t1 and t2 tables, write a query to avoid duplicates?  
   Assume, we have two fields/columns in the database as key and value

Select distinct(values) from t1 where values (select values from t2 where values not null)

1. Have t1 and t2 tables, write a query to avoid duplicates and count the values?

Assume, we have two fields/columns in the database as key and value

Select count( distinct(values)) from t1 where values (select values from t2 where values not null)

1. Tomcat server is running up, application page openings in few hosts and same application page not opened in few hosts, log file is plain…how to debug?  
     
   i) we need to ping the host where tomcat server is running, whether host is reachable or not.

ii) may be problem with dependent servers or applications

1. Mapreduce program to count words?

i) Write logic to parse the data which separates integers in Mapper class  
 use string tokenizer to separate the values  
ii) take a global variable **sum/count** which is initiated to zero.  
iii) Iterate the values with in the for loop and increment **count** inside the loop

1. How to achieve joins in hadoop?  
   Need to use distributed cache mechanism
2. How to sort values in the Mapper to Reducer?

Use secondary sory, so that values will be sorted from mapper to reducer.

1. How datanodes will signals namenode [is in HDFS?](http://www.aptibook.com/discuss-technical?uid=tech-hadoop13&question=What-is-a-heartbeat-in-HDFS?)

A data node sends heartbeat to Name node and task tracker will send its heart beat to job tracker. If the Name node or job tracker does not receive heart beat then they will decide that there is some problem in data node or task tracker is unable to perform the assigned task.