1. How will you get a hdfs file into a local directory?
2. How would you see the running application in yarn from the command line ?And how will you kill the application?
3. Say you have data of a website containing information of a logged in user ,one user may have multiple fields. But the number of fields per user may vary based on his actions.In that case which component of hadoop you will use to store the data?
4. Say you have one hbase table. Is it possible to create a hive table on top of it? it should not be manual data movement activity ,Any changes in hbase table should replicate in hive table.without any changes or data movement ?
5. Assume If data from external sources is getting populated in to hdfs in csv format on a daily basis,How would you handle it efficiently so that it can be processed by other applications and also reduce the data storage
6. There are 5000000 Records in one hive table and you have loaded it in spark -shell for development purposes. What would be the best practice to write code.Would you be processing 5000000 records in each line of code?
7. If there is a csv file present in hdfs location which has a header . while reading it in to spark, which property needs to be set
8. There is one csv file and you want to load it into the spark dataframe. You do not want spark to inferSchema for that csv file and you have a custom schema based on your requirement , How would you create a custom schema and assign it to dataframe ?
9. In hbase how to check whether tables exist or not ?
10. There is a dataframe df1, how would you cast all the columns of the dataframe to string , There should not be any hardcoded values?