Please find the below questions asked to me for JPMC interview

1)initially explained me about their implementations

2)asked me about the projects I worked on, overall and relevant experience

3)started asking me the questions on what I worked on

4)Sqoop

a)how sqoop gets connected to different databases?

b)have you used any specific delimiter in sqoop?

c)default mappers run in sqoop

d)how to import data if no primary key specified in the table?

e)what is split-by in sqoop?

f)if I have 10 different dates and I use split by date(--split-by date) and I am specifying  no of mappers 4 (--m 4 ) so how these dates are distributed across these 4 mappers, inner logic

5)Hive

a)difference between  internal and external table

b)serdes in hive

c)syntax for json serde, how to insert data to a json table from normal table(managed/external) and how schema is mapped?

d)performance tuning in hive(explained him about orc format and snappy compression)

e)while issuing query on orc table in hive, so will data be de-serialized that time? Once data is de-serialized how it will guarantee on better performance?

f)are you able to see the content of orc files?

g)Map Side Joins

6)Pig

a)how to filter out the data having(special character like enter,tabs) in pig,have you come across this situation?

b)given me a scenario that user while filling the data pressed ’ENTER’ many times so that the data is spread in new lines. How to process this data in pig so that  this data should come in single line? And no field should be deleted.

c)given me scenario that they have tab delimited data and this data contains tabs in between so how to process this data?

7) kafka

a)what are topics, consumers in kafka?

b)how consumer will know from which topic the data should be consumed?

c)once the data is consumed how to process the data?

d) how spark knows which consumer it should receive the data?Configuration?

e)how kafka know which the most recent data/or data is already consumed?(concept on offset in kafka)

8)Any experience on impala,hbase?

9)How to address small file problems in Hadoop? If you get (1kb data ,3 gb data etc ….) and you are not sure about the size of the data so how do you design your Hadoop system?

FirstRound(By Techlead from client):

===================================

1) Explain what you have done in your previous projects ?

2) What challenges did you face ? And how did you resolve?

3) One query is using a inner join for two tables, it is running all mappers and reducers quickly within first 2 or 3 mins.

But one reducer is running long for 3 hours and struggling to finish it. What is the problem here ? And how do you rectify ?

4) What are file formats and compressions ? which one you are using ? why ?

5) Two tables are cross joined and giving many duplicates ? How do you remove them in the result dataset ?

6) Why do we use Partitioning and bucketing ? which one to use when?

7) Performance tuning techniques in Hive ?

8) What is Mapside join? when to use it ?

9) What is Reduce side join ?

10) Did you work on Mapreduce ?

11) What are execution engines in hive?

12) Types of tables in hive ? Which one you use? why ?

13) What is left semi join ?

14) What is explain, what is analyze, what is indexing ?

SecondRound(By Client Manager) :

==============================

1) Explain what you have done in your previous projects and challenges ?

2) Performance techniques in hive ?

3) What are all lang's Spark supports ? scala,python,java. My team member is using scala ? why ? Is there any difference in performance for scala and java ?

4) Join techniques ?

5) Finally, he will explain about their current project work and will ask if you are willing to work in that kind of environment ?

**Tip :  Just be confident and well prepared for first Question i.e, explaining about your previous/current project and challenges you faced, how did you resolve them ?**

Pf below Key Bank interview Questions on Hadoop and SQL:

**Hadoop:**

1)      Mapper,Partitioner,Combiner,Reducer,Sort & shuffle phase.

2)      While running mapreduce program where mapper output will reside?

3)      Without reducer whether we can run Mapreduce program or not?

4)      Difference between Internal & external tables?

5)      Mapred modes in Hive(Strict & non strict)

6)      Which mapred mode used in Production environment and what are limitations when we use strict mode?

7)      How to enable TEZ engine in Hive

8)       Partitioning & Bucketing with example

9)      MSCK command in Hive

10)   2 to 3 questions on scenarios of Partitioning & Bucketing concepts

11)   File formats used in Hive

12)   Why we need to use ORC?

13)   What is Vectorization and which property need to set?

14)   Why only ORC file need to use for Vectorization?

15)   How many MR jobs will initiate when you will run Select \* from tablename;

16)   What is the signafance of Explode function in Hive?

17)   What is the default serde used for CSV

18)   What are the compression techniques you know other than LZO?

​19)    What Analyze command will do in hive?

20) How incremental loads will performs in Sqoop?

21)When we will use Pig and Hive?

23)Export Function in Sqoop?

**SQL:**

1)Joins

2)Sub queries

3)Correlated Sub queries with example

4)Group by and Having Clause.

5)Explain what is Rank,Dense rank,rownum

6)5 to 6 Queries on Joins and groupby having clauses.

7)About Isolation levels

8)What is cursor and Types of cursors

9)What coalesce() will do?

10)How to get top 6 salaries in a company table?

if hive.mapred.mode = strict.  
  
Strict mode is a setting in Hive that prevents users from issuing queries that could have unintended and undesirable effects.  
Setting the property hive.mapred.mode to strict disables three types of queries.  
  
First, queries on partitioned tables are not permitted unless they include a partition filter in the WHERE clause, limiting their scope.  
hive> SELECT DISTINCT(planner\_id) FROM fracture\_ins WHERE planner\_id=5;  
FAILED: Error in semantic analysis: No Partition Predicate Found for Alias "fracture\_ins" Table "fracture\_ins"  
  
Second type of restricted query are those with ORDER BY clauses, but no LIMIT clause.  
hive> SELECT \* FROM fracture\_ins WHERE hit\_date>2012 ORDER BY planner\_id;  
FAILED: Error in semantic analysis: line 1:56 In strict mode,limit must be specified if ORDER BY is present planner\_id  
  
Third and final type of query prevented is a Cartesian product.  
hive> SELECT \* FROM fracture\_act JOIN fracture\_ads WHERE fracture\_act.planner\_id = fracture\_ads.planner\_id;  
FAILED: Error in semantic analysis: In strict mode, cartesian product is not allowed. If you really want to perform the operation, set hive.mapred.mode=nonstrict+