* 1.Project flow
* 2.How to read a csv using Scala and spark
* 3. If joins are performed,how will you check whether it is working fine or not?
* 4. If u don't have pk, how will you achieve parallelism?
* 5. If you can't find a split by col, how will you do the import?
* 6.Word count program
* 7.Recursion program in Scala
* 8.String reverse program with out using built-in func
* 9. Types of tables in hive, on what tables you have worked in ur prjct
* 10. If I have 500 files with 500MB how will you move to hdfs
* 11.If I have 100Mb file how many mappers u will use in sqoop
* 12.Spark submit command
* 13. How will u execute spark submit command
* 14. Have u executed spark submit?
* 15. How to impose schema to rdd? What is the difference between case class and struct type
* 16.Types of joins in hive apart from normal joins
* 17.Internal working of joins
* 18. Difference between map and flatmap
* 19.Difference between map join and bucket map join
* 20. Optimization techniques in hive
* 21. vectorization is not applicable for csv\*\*\*
* 22. Partitioning,bucketing
* 23. Second highest sal deptwise
* 24. And except first highest he want all the highest sal in Dept deptwise
* 25.Have you worked on rdd and DF,dataset
* 26. Higher order func
* 27.If u write a function in one line,what is that function called
* 28.Higher order func sample code
* 29.Tek systems-DXC technologies
* 30.Project flow
* 31.HOF, example
* 32.What is casting, when will go explicit casting
* 33.In ur project,tell me a scenario where u have done casting
* 34.If u provide elements in b/w paranthesis,what is this called in Scala?
* 35.How to access elements in tuple
* 36. When we go for tuple?
* 37.What is the limitation of tuple
* 38.What is the difference between pure func and normal func
* 39.What are the side effects in pure func explain with example
* 40.everything in Scala is expression? Justify
* 41.Is for loop is an expression? How explain
* 42.What is the o/p of for with yield
* 43.How to convert the elements in df to list
* 44.Objects hierarchy in Scala(data type)
* 45.How will you optimize the queries in hive?
* 46.What are the types of index in hive?
* 47.( Compared with sql)
* 48.Can we truncate the data partition wise?
* 49.If the format in file differs the format what u want to load,how will u insert the data in to table
* 50.Will u specify partitioned column in insert stmnt
* 51.Optimizations in hive
* 52.Coalesce and repartition
* 53.In which scenario u go for coalesce is it necessary to go for coalesce
* 54.(Based on some property we can append the files in hive it seems ) in this scenario no need of coalesce then how?
* 55.How to join multiple DF's
* 56.Providing Schema to RDD
* 57.What is Nil,Null?
* 58.Do you know anything or else?
* 59.When a query is executed,explain what happens internally? In hive and sparksql
* 60.What are transformations and actions?
* 61.Will Transformations execute as Soon it is triggered?(no lazy eval).
* 62.Where the transformations are stored in Rdd?
* 63.Rdd is fault tolerance,how it is achieved?
* 64.Difference between Rdd lineage and DAG?
* 65.In query flow in sparksql sql, how will u get metadata ?
* 66.What is catalyst and tungsten?
* 67.What is orc format? Why it is used?
* 68.Instead of orc format I can create index for the columns I need , justify?

TCS:

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

* 69.What is monad ?
* 70.Replication factor? Default where it is set?
* 71.What formats u have worked in project?

72. what is the difference between map and map partition. explain with example and where to use when?

2. scenario - suppose if an rdd is deleted in spark job how do we can recover it and what is the backend mechanism of recovering rdd's.

73.how do you connect to your cluster using data nodes or edge nodes and what is reason of choosing between the both?

74. have you ever received an error "spaceout" in your datanode?

75. how do you allocate buffer memory to your datanode?

76.how much buffer space have you allocated to your map task and reduce task in your datanode?

77. how do you achieve broadcast join automatically without out doing it manually? and how do you setup your driver program to detect where broadcast join can be good to use and how do you automate the process?

78. how do you acheive in memory caache?

scenario : imagine you are working on cluster and already have cache your rdd and got the output stored in cache now i want to clear the memory space and use that space for caching another rdd? how to achieve this?

79.what are the packages you have worked in scala name the package you have imported in your current project?

80. what modules you have worked in scala and name the module which you have worked till date?

81.Kafka - scenario : suppose your producer is producing more then your consumer can consume , how will you deal such situation and what are your preventive measures to stop data loss?

82. how do you achieve " re-balancing in Kafka and in what way it is use useful?

83. Kafka scenario : suppose producer is writing the data in CSV format and in structure data then how will the consumer will come to know what schema the data is coming in and how to specify and where to specify the schema?

84. how do you manage you offsets?

85 . Kafka : scenario : suppose consumer a has read 10 offsets from the topics and it got failed then how consumer b will pick up offsets and how does it stores the data and what is the mechanism we need configure to achieve this.

86. Hive : Scenario: Imagine we have 2 tables A and B.

B is the master table and A is the table which receives the updates of certain information

so i want to update table B using the latest updated columns based up on the id how do we achieve that and what is the exact query we use?

87. What is use of Row-index and in which scenarios have you used it in hive?

88. what do you know about Ntile.

89. Spark - Scenario : Suppose i m running 10 sql jobs which generally take 10 mins to complete, but one it took 1 hour to complete if this is the case how to you report this error and how will you debug your code and provide a solution for this.

90. Spark submit

91.Spark initialization with main method

92.Empid 2011 2012 2013

1. 100 200 300

2 300 400 500

Add all salaries row wise.. And generate two columns one is Empid and other is new column which generates total sal

93.Only one column in table. Need two columns in output. New column should contain if starts with letter T then truck, starts with C then car other than this unknown should print...

Vehicle no

T 01234

C 01345

X 14554

A 09835

Output shud be:-

Vehicleno newcolumn

T012345 Truck

C01345. Car

X13455. Unknown

A34556. Unknown

94.Difference between RDD dataframe and dataset

95. Though we can write spark application both with Scala and python.There is one drawback with python? What is that?

***Tek Systems(DXC)***

***96.***

|  |  |
| --- | --- |
| Emp\_Id | Dept\_No |
| 1 | 10 |
| 2 | 20 |
| 3 | 30 |
| 4 | 40 |
| 5 | 50 |
| 6 | 10 |
| 7 | 20 |
| 8 | 10 |
| 9 | 30 |
| 10 | 10 |

Here I want count of Dept\_No = 10 ?

97.

|  |  |  |  |
| --- | --- | --- | --- |
| Id | Name | Sal | Gender |
| 1 | Abc | - | M |
| 2 | Xyz | - | F |
| 3 | Pqr | - | M |
| 4 | Kiran | - | M |
| 5 | Kiran | - | F |

Here I want **Name** of which satisfies bothgender male and female?

***CAPGEMINI***

98.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id | Sal1 | Sal2 | Sal3 | Sal4 |
| 1 | 100 | 200 | 300 | 400 |
| 2 | 50 | 60 | 70 | 80 |

|  |  |
| --- | --- |
| Id | Total sal |
| 1 | 1000 |

Requirement is should be show like

Total sal= (Sal1+Sal2+Sal3+Sal4)

99. Using Dataframe

|  |
| --- |
| Vehicle id |
| T1234 |
| C890 |
| Any thing else |

Requirement is adding a new column

like if

starts with t -> truck

starts with c-> Car

starts with other -> Unknown

should be shown in the form of

|  |  |
| --- | --- |
| Vehicle id | New\_col\_name |
| T1234 | TRUCK |
| C890 | CAR |
| OTHER | UNKNOWN |

100s .from below table I want 2 dataframes even numbers should be stored in 1 dataset and Odd numbers should be stored in another dataset

|  |
| --- |
| Id |
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

101.What is sparkcontext