**HIVE INTERVIEW QUESTIONS**

What are the different types of tables in hive? - celebel technologies

You have a table in Hive is some partition column. If we delete partition column what will happen to the directories created? - Emid Technologies

If you don't specify the external keyword while creating external table which location will it take.write the location path. -Emid Technologies

 Write a HYPE ddl statement to create a partitioned table -IQVIA

Explain analytical functions in HQL  -IQVIA

I have a hive table and I am running a query over my Hive table and I am getting a outOfMemory exception, how to avoid it ? -fractal analytics

 What table have you use in your project ? -fractal analytics

If you are joining multiple tables based on some key column and some key is coming multiple times in a row then whether it will create any problem? -fractal analytics

 Have you worked on UDF's what are the jar dependencies you have added in your UDFs? Difference between UDF and UDAF -Systech solutions

Have you used SCD2 in hive? -Systech solutions

  I want to pass a table name in HIVEQL, parameterised, how to do it? -searce cosourcing

concat\_ws and explode function in hive ? -searce cosourcing

 Wright sqoop code that you are using in your project for exporting data to HIVE? -HSBC

Hive table and asked to write output for inner outer joins. - HSBC

Suppose there are 2 tables one with a 0 bucket and another with 3 buckets. Now there is another table that I join it with 0 bucket table and in another case with 3 bucket table. Which join will be faster? -YES BANK

Suppose there is huge dimension and huge fact table which is joined on the basis of the customer column.one of the joining  keys is contained by the 99% of the date. (Say customer I'd =1).my reducer is stuck at 99%? How would you resolve this problem in hive? -YES BANK

Explain the internals of ORC file ? -YES BANK

 If you have timestamp value in one column in hive ? can you calculate the difference of two consecutive rows in timestamp in another column? -YES BANK

What is the difference between SCD2 and SCD3 ? -TO THE NEW

If you have collection in Hive how do you display the each collection item in Hive in a separate rows? -TO THE NEW

What is the difference between Snappy and gzip compression technique ? -TO THE NEW

What is HCcatalog? -TO THE NEW

How will you do Pivoting in hive ? -TO THE NEW

Spark program has 3 input file, data in these 3 files are numbers from 1 to 1 billion number, i.e. (1,3,4,2,6,7,8,9,5…. till 1 billion)

Now write missing numbers logic in optimized way in any logic [Missing numbers should be saved in DataFrame].  - Tavant Technologies

There is a legacy MapReduce program, now there is an enhancement to do, but instead of MapReduce, you need to do it in Spark, But you shouldn't rewrite MapReduce code to Spark code, in this process, can you write your enhancement code in Spark & make it invocable in MapReduce code? If yes, how is it possible? & how does it work?  - Tavant Technologies

When will we go for external table in hive? Explain a scenario where you go for external table -HCL

You have an order table which is of 50 GB size and will daily data of 10 GB. New data can be an update to existing data or will be an insert in case of new records. How will you design the table and HIVE script for optimal performance.  -HCL

 What are the optimisations we can do in Hive ? -Citi bank

Write 3 GIT commands which you are using daily. -HCL

How is your project handling incremental load when the requirement is to append data to the target table which is HIVE? -HCL

Complex data types supported in HIVE -HCL

Target table [Country\_id + Year as key]

Country\_id,Year,Total\_Population

IND,2020,5000

PAK,2020,4000

SOURCE [source table will be of format TableName\_YYYYMMDD]

T1\_20201212

COUNTRY\_ID,COUNT,Type

IND,2000,BIRTH

IND,200,DEATH

PAK,200,BIRTH

SRI,2000,BIRTH

\*Get the year column from source hive table name

\*update the target table based on source data

For the above data , target should be like

Country\_id,Year,Total\_Population

IND,2020,5800 (5000+(2000-200))

PAK,2020,4200  (4000+(2000-0))

SRI,2020,2000

-WALMART LABS INDIA

We already have a feature partitioning then, Why we need bucketing in Hive. - WALMART LABS INDIA

Optimisation techniques used in the project. - WALMART LABS INDIA

We have a target table or details- with > 50 GB size and daily source data is >5gb. How will you design the target table (partition columns) and how will be the job design.

1) We have to identify new records and insert.

2) Already existing records should be updated because there an be status change. - WALMART LABS INDIA

How to delete records in hive? - Sapient

What is the difference between the internal and external tables? - Lumiq

What is CDC? - Lumiq

How to fetch updated records in hive? - Lumiq

What is distributed cache? - Apple Tech Consultant

What is ORC, parquet file format in Hive? - Apple Tech Consultant

Suppose I delete the external table by mistake. Now I want to recover that data? How I recover the deleted data of the external table? - Apple Tech Consultant

Which are the 4 different ways to get unique records? - Apple Tech Consultant

What are the optimisation techniques used for Hive in your project? - Infosys

Explain how bucketing and partitioning used in your project. - Infosys

What is the difference between partitioning and bucketing in Hive? - Capgemini

What is the internal and external table in hive and the difference between them? - Capgemini

Example of static and dynamic partitioning and the difference between them in hive. - Capgemini

Give high level overview of Hive? - ImaginTechServices

What all issues you have faced while creating bucket? - ImaginTechServices

Explain static and dynamic partition. - Imagine Tech Services

How we modify the datatype of a column in Hive? - Imagine Tech Services

What is Schema evolution in Hadoop?

There is a Hive table with 10 columns and the files are sitting in Hive HDFS path with 10 columns values. Now suddenly the source team had added one additional column and sending the files? What will happen in this case?

I want to load files of approx. 1TB of size to a Hive table and I’m more inclined to run aggregation queries against few columns. How will you design the solution in a way that your query performance is optimized? - CTS

 We have two tables policy and insurance. both tables contain policy\_ id column. We need to group all the insurance done by a particular customer then join it with the policy table. The final table should contain all the policy\_id irrespective of whether the insurance has been done. How should the join be done? - IBM

 Give examples of where you have created external table and internal table.- IBM

 Explain when do we need to create partition and when bucket.- IBM

 Can we delete any partition?- IBM

 Explain indexing. - Alliance Data

 Can we change the location of the already created table in hive?- Alliance Data

 On which column do  we build partition.- Alliance Data

 We want to give read permission on particular columns of a table. How can we do that?- Alliance Data

 What is Hive BucketedMap-Join? - HCL

 Maximum number of buckets can be created in Hive? - HCL

How will you join two tables in Hive having 100GB of data ? - CTS

What are the different optimisation technique available in Hive?- CTS

 Difference between Hive and Impala?- CTS

 If one of your team member has loaded the data in one table in Impala, you can see the table but not the data then?- CTS

Can you able to access Impala table in Hive?- CTS

 How to delete and replace work in Hive?- CTS

 How to update data in Hive/ Impala?- CTS

 If you have loaded the data in Hive but your team member is not able to see. All the privileges has been provided to the table. - CTS

 Which version of Hive, Impala, sqoop, Hadoop you are using?- CTS

 How will you delete records from internal/manage tables?- CTS

 What are the various execution engines in hive?

How to set properties in Hive?

What is vectorisation?

How to set parallel parameters in hive?

Explain statements in Hive?

Partitioning and bucketing in Hive and use cases?

Difference between cluster by and distributed by in Hive?

Lateral view in Hive?

What are different types of joins in Hive? Explain with scenarios?

- Paypal

 Define the difference between Hive and HBase?

Explain Hive architecture?

What is the use of Thrift server in Hive?

What is the use of Metastore in Hive?

How to do performance tuning in Hive?

Partitions and Bucketing in Hive. -- Create bucket table, then query and browse.

RANK() vs DENSE\_RANK() functions in Hive?

How does Hive decide number of mapreduce jobs it runs?

Advantages vs Disadvantages of using Derby as hive metastore?

- American Express

 What is external table in Hive? - EY

 Difference between partitioning and bucketing? - EY

 How to delete a record in Hive? - Iris software

 How to find out the latest transaction in the table for the given transaction\_ id and transection\_time? - Iris software

 What is partitioning and bucketing in Hive? - City Bank

 What is external and internal tables in Hive?- City Bank

 What is the benefit of using internal tables in hive? - City Bank

How would you declare a Hive table as internal or external? - City Bank

What is Hive partition and bucketing? When to use which? - Clairvoyant

Explain Map side in Hive? - Reliance Jio

What is difference between partitioning and bucketing in hive with example. - Cognizant

What is SerDe in Hive? - Cognizant

What are Sparse files in Hive? - IBM

What is the difference between ORC and Parquet file format in Hive? - IBM

Which file format should I choose under what circumstances? - IBM

I get data on daily basis, 3 files for insert, update and delete records respectively. I want to load them daily into my table. Architect this requirement, how you will do it using hive?  - IBM

Write logic in Hive to extract year from timestamp? - Wipro

How to test hive in production? - Wipro

How to use vectorization in hive? - Virtusa

How to implement cdc (change data capture) in Hive? - Sapient

Hive complex data types use cases.

How vectorization is helpful? - Legato

Calculate customer spend analyzer result over last 3 months for any bank for analytics purpose using purchase order in hive?

Hints:

Input:

order\_id category purchase\_amount purchase\_date

1 grocery 1400.00 2020-01-03

2 travel 500.00 2020-01-03

3 medicine 700.00 2020-01-03

4 miscellaneous 600.00 2020-02-14

5 grocery 1600.00 2020-02-05

6 grocery 1200.00 2020-03-03

7 medicine 700.00 2020-04-15

output: if your selected date is 2020-03-03

Category some %\_spend

grocery 4200.00 70%

travel 500.00 8.33

medicine 700.00 11.67

miscellaneous 600.00 10

- Wells Fargo

In India student were appeared for MBA exam and now rank them in percentile according to there score they got in exam?

- Wells Fargo\*\*\*

How to get how many case were compliant and how many days it took to reach compliant or non compliant between a audio history table and if sla\_date\_count=< 20 then met\_sla yes Else No?… - Wells Fargo\*\*\*

What are Map side joins in hive? - Synechron

Can we change the warehouse of where the data is getting stored in Hive? - Cognizant

We have an internal table with location.The table is dropped, what will happen? - Cognizant

What will happen if Hive metastore is down but hiveserver2 is running? - Cloudera

Coalesce/ repartition. - Altimetric

Converting external table to managed table command. - Altimetric

What is Hive to spark command? - Altimetric

What kind of optimizations used in Hive? - Altimetric

Why ORC Files used in hive? - Altimetric

Write a program which will find numbers which are divisible by 7 but not a multiple of 5 between 2000 and 3200 (both included). The numbers obtained should be printed in a comma-separated sequence on a single line. - Altimetric

How will you handle corrupted file in Hadoop? - Altimetric

What is SerDe in Hive? - Altimetric

How do you selected partitioning or bucketing on a dataset? - Altimetric

What is default SerDe in Hive?  - Altimetric

Can a Hive table have both static Dynamic partitions? - Altimetric

What msck command will do in hive? - Emid technologies

How to implement bucketing in hive, decide the number of buckets, and how it helps in the hive in execution? - Happiest\_Minds

Explain different types of optimization techniques in Hive? - TCS

Suppose you have a Hive query which normally takes 30 mins to complete everyday, but from last few days you are seeing the query is taking longer time to complete.

How will you troubleshoot the issue? - TCS

What optimization techniques are used in your project? - Iris software

What are views in hive and how it is beneficial? - Iris software

Suppose you have 1 table with 1 column and that having 7 distinct values and 3 null values, Now I'm running 3 queries, tell me the output of these 3 queries. Select count(\*) from tabke\_name, Select count (1) from table\_name, Select count(column\_name) from table\_name  - Iris software

Suppose I am having 1 table which hold student library data. - Iris software

Table structure. - Iris software

What all performance tuning techniques used in Hive? - Capgemini

How partitioning and bucketing works internally? - Capgemini

Can we point two external tables to a single file? - Capgemini

Can we partition the data on Timestamp file? - Capgemini

Write a query for Avg salary for every dept? - Capgemini

Difference between different file formats like ORC,RC,AVRA, Parquet. - Capgemini

Describe what all performance tuning techniques are used in your project? - ITC

Why ORC is best to use and how the data is stored in hdfs? - ITC

But default which file format data is stored? - ITC

Write a query to get dept,state,designation,no\_of\_employess, sum(salary) for each dept? - ITC

What is Static/Dynamic partitioning? - ITC

Difference between hive and beeline? - ITC

How to remove special characters from full table of hive? - CTS

Write a query to get the list from one column into multiple rows.  - CTS

How to get the running sum of salary for an employee over each department? - CTS

How bucketing works internally, how it is decided that which row will go to which bucket? - CTS

Explain the architecture of hive? - Impetus

What is transactional mode in hive? - Impetus

Difference between concat/concat\_w. - Impetus

What is the order sequence of hive- site.xml,txt file, command line arguments? - Impetus

Can we update an Hive table if yes, what prerequisite are there to update a table? - Impetus

Difference between external and manage table? - Xebia

You have two columns with value 1, Array["A","B","C"]. Now convert this one row into three rows

1,”A”

1,”B”

1,”C”

- Xebia

Write a sqoop command and explain how to import data without any unique column? - TCS

What is Metastore in hive? - TCS

What is Hive variable? - Legato

What classes are used by hive to read / write data from hdfs? - lowes

What is Serde in hive? How does it work internally(internal implementation) - IBM

If I delete an external table and then again create, will it be created successfully with all the data I need? Or is there any extra steps I need to do? - IBM

How does views and indexes work internally in hive? - IBM

How to know last value an incremental append in sqoop? - IBM

What is catalyst optimizer, how does it work internally? - IBM

1st round will be a technical coding round consisting of 3 questions in 2 hrs. If you complete 2  then you are eligible for next F2F round. - Sapient

Different file formats used in project? Parquet advantages? - Sapient

Which spark optimization techniques used in project? - Sapient

End to End working of Spark job, from submission to execution? How Yarn scheduler works? - Sapient

Can we have partition for indexes in hive? - Sapient

How to load the data from HDFS to Hive tables? - Verizon

How do you handle incremental in sqoop/ Hive and how to handle insert/update? - Verizon

Hive partitioning types of partitioning. - Verizon

DDL functions of Hive? - CTS

What is partition and bucketing? - CTS

Why manage table in our project 1? - CTS

Types of hive Tables and where you used this table in your project. - CTS

If the given data looks like

12,34,56

23,56,86

45,67,56,87

56,77,66

only one row having one extra column, how do you process it? - CTS

What are the hive process you are doing in your project.explain? - CTS

What are the requirements for hive table creation and how you are creating in your project? - CTS

How to automate that hive process and requirement for the automation? - CTS

Explain the requirements in Details. - CTS

What type of tables you are using in your project? - CTS

If you are using partition and bucketing in your project. - CTS

What is your default partition using in your project? - CTS

Is it possible to delete and update in hive Table? Have you used in your project? - CTS

What are the optimisation techniques used in hive? - CTS

Explain bucketing and partitioning? In which scenarios you use both in your project? - CTS

Write the query to create bucketing table. - CTS

How you will insert values to the hive table from one table using case condition? - Wipro

Syntax of map side join and why we go for map side join.. - Wipro

What is the max size of map side join small table? - Wipro

On sale directory if you create internal and external table, if you delete external can you see? - Wipro

Internal table details - Wipro

What is Hive metastore, where it is saved in prod cluster? - Wipro

Hive external table use. - Wipro

Can we use hive for OLTP systems? - TCS

What are hive table types and what you use in your project? - TCS

How to change Warehouse directory? - Legato

What is serde in hive? - Legato

What is the limitation of Derby database? - Lowes

What are complex data types in hive? - Lowes

Pre requisites and how to perform  a SMBJ join? - Lowes

Advantages of advanced joins in hive? - Lowes

In Hive, if you will do "select \* from table" then how many reducers it will create? - Alliance Data

If you do rdd.coalesce(0) then what will happen? - Alliance Data

How will you remove duplicates from dataset when a subset of 3 columns are common between the records? - Alliance Data

I have duplicates in hive table or a dataframe, how you find out duplicate records from that table/dataframe? - Alliance Data

How to update records in Hive? - TCS

In daily use case why I should prefer to use ORC file format over Parquet? - TCS

What kind of thing we can do with ORC that is not possible with Parquet? - TCS

Can I use ORC if I need to update the table? - TCS

When can I use partitioning & When I can use bucketing? - TCS

Hive table with differences. - HCL

What are different hive performance improvement techniques used in your project? - HCL

Maximum no of partitions we can do on a table in hive. - HCL

What happens when we reduce the number of buckets from 10 to 6 on hive table? - HCL

What are various file formats supported in hive? - HCL

explain about indexing and bucketing in Hive? - HCL

 Read json file using Spark, convert the rdd to dataframe and give schema to it, apply Spark SQL queries on it? - HCL

 Which hive built-in functions you have used and explain hive user defined you used in your project? - HCL

 How to update Hive table to low-level? - HCL

 Hive query to find 23rd salary from emp table. - Synechron

 Hive partitioning bucketing. - Tetrasoft

 Hive optimization techniques. - Tetrasoft

 Hive query to get 2nd highest sum of salary of department ( given emp and dept tables). - Tetrasoft

 Hive internal and external tables. - Tetrasoft

 When do we use partitioning and bucketing? - L&T infotech

 What is the difference between Derby and any other RDBMS metastore? - L&T infotech

 Difference between schema on read & schema on write? - L&T infotech

 How is Hive different from traditional RDBMS? - L&T infotech

 Difference between Hive and Impala? - Apple

 Difference between Parquet and avro? - Apple

 Difference between semi join and anti-join? - Apple

Where internal Hive table data will store in hdfs? - Apple

 If we drop the partition, will be able to access the data? - Lumiq.ai

If we drop the Hive table, will the view be assessable? - Lumiq.ai

 Difference between partitioning and bucketing? - Lumiq.ai

 Difference between internal and external table? - Lumiq.ai

 Different Hive optimization techniques? - Lumiq.ai

 How do you remove duplicates without using group by function? - Alliance Data

 Select \* from employee where salary>4000 group by dept order by salary limit 10;

In above what is the order of execution of clauses. - Alliance Data

 Select \* from employee where salary>4000 group by dept order by salary limit 10;

In above what is the order of execution of clauses. - Alliance Data

 How do you remove duplicates without using group by function? - Alliance Data

 HiveQL vs SQL- Order by, Sort by, partitions, joins, limitations, storage(warehouse), Metastore. - HCL America

 Hive external table vs internal table. - HCL America

 Sqoop CLI commands -copyFrom, -copyTo, -put, -import, -export, -job, -merge, -exec, -list, append, fileformat, compress, password file, parallel.  - HCL America

Presto Architecture, Components, functions, advantages for data analysis over Hive? - HCL America

Internal vs external table? - Herman

 Write a query to create joins? - Herman

 Hive job is running slow how to troubleshoot? - Herman

 How to copy Hive tables from one cluster to another? - Herman

 Explain Hive architecture? - Modak Analytics

 What is role of Hive meta store? - Modak Analytics

Explain Hive query life cycle? - Modak Analytics

How to do load balancing in hive? - Modak Analytics

What is difference between partitioning and bucketing? -Mindtree

What is use of strict mode in hive? -Altimetric

In hive, how do you move data from Parquet table to orc file format? - Altimetric

How to check for long running jobs in hive and how will you handle it? - TCS

What is SCD?and its types. - TCS

How you handle small files in hive? - TCS

Which tables you have created in your project? -Mindtree

How to update batch data in hive table?-Mindtree

Which join technique you have used in your project?-Mindtree

What happens if multiple hive job take data to process at a time from a shared folder? -Mindtree

What is SCD?-Mindtree

What is acid property?-Mindtree

Write a query to find the name of each employee having same salary? - L&T Infotech

Write a query to load a text file in hive table and I want to store as parquet file. - L&T Infotech

How can we transactional load in hive table?- L&T Infotech

How will you handle corrupt data in Hive?- L&T Infotech

Which hive Table you have created in hive in your project? - L&T Infotech

I want to keep all table structure but I want to delete records permanently which should we use? - L&T Infotech

Aggregate,Rank,Row number functions in Hive SQL. - Optum

Do you know about blocking in SQL? - Optum

Advantages of Hive External tables.explain your use case? - Land O Lakes

Read a csv file in hive and ignore first 2 records. - TCS

What is partitioning and bracketing? Give example. - TCS

Difference between partitioning and bucketing? - Wipro

What is Dynamic and Static partition? - Wipro

On day 1, you received a file t1.txt which has certain records. - Legato

Explain lateral view explode function? - Virtusa

Incremental load data? How to find the top 3rd salary in the employee table. - Accenture

Diff b/w explode vs lateral view.

sorting techniques in Hive

partition vs bucketing? - Neustar Inc

How we can handle the size of map and reducers If we face OOM issue?  - Bold Technologies

How to kill the running application of yarn via command line? - Bold Technologies

What is the difference between partitioning and bucketing? - Bold Technologies

Difference between static and dynamic partitioning?  - Bold Technologies

When to apply partitioning and when to apply bucketing basically with column we need to select to apply these conditions? - Bold Technologies

SCALA INTERVIEW QUESTIONS

What is future class in Scala? - Legato

What is diamond problem ? write a code to solve it in skeler ? - Searce Cosourcing

Write a program to find the prime and composite number in Scala? - Yes Bank

What do you mean by functional programming in Scala ? - Yes Bank

Let us say that there is one function where in you pass the date and you have to check the format of date as 'mm-dd-yyyy' also if the use pass invalid date, that has to be detected too. How will you create this function? - TO THE NEW

Write a program to find the missing element from the list1 where you have list1 and list2

List1 (1,2,4,6,7)

List2(3,5) - TO THE NEW

What is diamond problem in Scala and how this can be resolved ? -  TAVANT

Given the below dataset, write a scala program to remove the duplicate records from it based on name and age.

employee.txt

Rajesh 21 London

Suresh 28 California

Sam 26 Delhi

Rajesh 21 Gurgaon

Manish 29 Bangalur

-  Sapient

Write a word count program in Scala ? -  Accenture

What is function currying in Scala? - Capgemini

Why Scala over Java? -  IBM

Why singleton object? -  IBM

What are companion class? - IBM

What are closures.Give with example? - IBM

Explain case class? - IBM

Explain yield with example? - IBM

What is the advantage of using higher order function Scala? - Sapient

Why there is no static keyword in Scala? - Reliance Jio

Suppose we have list given below, we need to find what is the maximum length where it is following the decreasing order.

10,9,7,6,5,4,3,2,1,8

 - Celebel technologies

We have string s="my name is manikanta", what will be the output when we apply map and flatmap? - Celebel technologies

Difference between apply method and un apply method in Scala? - Celebel Technologies

What is function currying in Scala?  - Wipro

What is the difference between closures and curring functions in Scala? - HCL

What is auxiliary constructor in Scala and why is it used?- HCL

Explain tail recursive function?- HCL

Use of case classes. - Altimetric

What are monads? What is the use of monads? - Capgemini

What is a companion object?Why do we need company object? - Synechron

What is Scallop in Scala? - Synechron

What is the difference between Map and Map partition? - Accenture

From the sample data given below, remove duplicates on the combination of Name and Age and print results.

Suresh 28 California

Sam 26 Delhi

Rajesh 21 Gurgaon

Manish 29 Bengaluru

- Sapient

What is case class and why it is used? - Capgemini

What is tuple, explain with a use case?- Capgemini

What is Higher order functions with query?- Capgemini

What is currying? Explain with example?- Capgemini

What is closure functions?- Capgemini

What is singleton object and companion object?- Capgemini

What is case class and what it is used and beneficial? - CTS

What is companion object?- CTS

 - Please do not use high level API/Framework like pandas/spark-sql etc.

 - Solve this problem by using simple data structures given in a programming language.

 - Please try to optimise the solution for efficiency in terms of space and time.

Given Dataset

Name age location

Rajesh 21 LondonWhat is the use of struct keyword and which case scenario it is used?- CTS

What is singleton object?- CTS

What is closure function, explain with example? - TCS

What are Higher order functions?- TCS

What are partial order function?- TCS

Explain the working of annotiate with example?- TCS

What is packages and how it is used in real time scenario. Why usage of package are beneficial in projects?- TCS

What are the features of Scala?- TCS

What is a trait in Scala? - IBM

What is the difference between Java if-else and Scala if- else?- IBM

Read an Avro file using scala with its schema file?- IBM

What are case classes?

What are singleton objects/ companion classes?- IBM

Difference between MAP and SET?- IBM

What is higher- order function in Scala?- TCS

What is the diamond problem in Scala?- TCS

How we can keep multiple inheritance in scala?- TCS

What is case class? - Synechron

Scala program for factorial. -Wipro

Scala case class.-Wipro

About singleton in Scala.-Wipro

Traits and monad in Scala.-Wipro

Use of yield in Scala.-Wipro

Scala traits and example. - Tetrasoft

Palindrome program in Scala. - Tetrasoft

Tail recursion in Scala. - Tetrasoft

Short data without using sort method in Scala. - Tetrasoft

How do you define constructor in Scala? - Knoldus software

What is singleton object in Scala?- Knoldus software

Is there any static keyword in Scala? if not how can you achieve static methods in Scala? - Knoldus software

What is trait in Scala. explain what can be achieved with trait.- Knoldus software

Explain the scenario where you need only one object for a class.- Knoldus software

Explain how Scala is different than Java. What are the differences.- Knoldus software

Difference between map and flatmap? - L&T Infotech

Explain lazy evaluation with an example?- L&T Infotech

Where are RDDs immutable?- L&T Infotech

What are all the higher order functions you have used?- L&T Infotech

Why scala is prioritized than Java? - Securonix

What is the trait in Scala?- Securonix

How to inherit traits and if theirs are multiple traits then how you will extend them. - Securonix

All available collections in scala.- Securonix

If you have immutable map, then how will you had elements in it.- Securonix

What is a singleton. why we use singleton class?- Securonix

Why case class generate setters, getters method and why we don't use new keyword to create object?  - Synechron

How can we find a element from a list in Scala?- Synechron

What is higher order function?- Synechron

How we achieve concurrency in Scala?- Synechron

What is case class?- L&T Infotech

What is curry?- L&T Infotech

What is tail recursive?example?- L&T Infotech

What is main difference between Java and scala?- L&T Infotech

You have to arrays. - Virtusa

How hashmap stores the data internally. explain. - Virtusa

What is a diamond problem? - Kreuzwerker

What are closures and implicit variables? - Kreuzwerker

What is future class in Scala? - Legato

Difference between fold by and fold by write in Scala? - Legato

Have you work on multithreading concept in Scala?- Legato

How do you fix production issues and tell me few issues which you have fixed in your project - Legato

Can we have non abstract variable in trade?

show the syntax of option with example?

Explain the apply method working?

explain fold left, I have an least (1,2,3,4,5) multiply each element using foldleft?

Hadoop Interview Questions

You might be aware of the web services used in your big data project. Can you name some of them which are commonly used? - Infosys

The questions about your project and its architecture. Be prepared with the problems you faced and what are the steps you followed to solve the issues. - Sapient

How to transfer data from one cluster to another cluster? -Infosys

If any job in production fails, How will you start debugging i.e what is the debugging flow to find the error. - Accenture

What is difference between client and cluster mode in Hadoop?- Accenture

Which file formats you have worked on in your project and what is the difference between all file formats? - Accenture

What is the cluster configuration in your project?- Accenture

Explain your project with an architecture diagram. - Capgemini

Size of cluster, RAM and how much the data getting processed daily in your project. - CTS

Difference between RDBMS and Hadoop.- CTS

What kind of semi- structured data you have work on.- CTS

You have 2 separate file one with 10 MB and another with 100 MB of data. How the data will get stored from 64MB point of view. - CTS

How will you check error in HDFS/ If in case any error log gets generated?- CTS

How will you rate yourself out of 5 in -- Linux, Hadoop, JAVA, SQL, Networking - Cloudera

Can there be two active name nodes at a time? (mention splitbrain and fencing)- Cloudera

What do you have to check for HA? (Like what should be working and all) (JN, ZK, NN) - Cloudera

I am trying to reach google.com but I am not able to reach, what thing I should check? - Cloudera

I am trying to reach https://website, I am getting certification error, what can be the issue? - Cloudera

I have multiple files with same name and same size, how can I differentiate it? - Cloudera

What is TCP/ IP protocol?- Cloudera

How can you check the logs for a currently running services? - Cloudera

There is a file with 1 lakh records having salaries of employees with location. Read the file and tell the average salary of employees based on their location. Then transfer the file in HDFS on the basis of location. - City Bank

Explain the toughest challenge you faced during your experience so far? (I explain the problem I faced while data node addition where service principle was not getting deleted automatically by CM) - Cloudera

What is your experience with an issue in the cluster? - Cognizant

How to define block size in Hadoop .. based on cluster or user's choice to configure?

Hadoop 2.x if the job should fail how to see log status in YARN.

How much RAM it will allocate for JVM in Hadoop cluster

 - Legato

Explain how authentication and authorization works in Hadoop ecosystem.- Cloudera

How to control access in hdfs data?- Cloudera

Explain how to setup Namenode HA?- Cloudera

Why do we need to choose 3 journal nodes at least in a HA setup and what will happen if 2 out of 3 journal nodes goes down?  - Cloudera

What will happen if zookeeper goes down in a cluster and how to find where zookeeper servers are running? (you don’t have access to the console) - Cloudera

How do you monitor cluster health/issues on daily basis? (Basically they were interested to know if we don't use any products like ganglia, fortunately in my last project I developed a tool in bash using Cloudera manager API, I explained them in detail) - Cloudera

How to copy hdfs data from one cluster to another?- Cloudera

Which Hadoop distribution you have used, Cloudera or hortonworks? - Cloudera

What is namenode? what happens if namenode fails. - NICE SOFTWARE SOLUTIONS

What is data node? what happens if data node fails.- NICE SOFTWARE SOLUTIONS

If I have 9 nodes in a Hadoop cluster and all of these nodes are full, if I lose 1 node out of it would my data still be available? Now in the same scenario, if I lose 7 nodes and only 2 nodes, will I still have all the data available. - NICE SOFTWARE SOLUTIONS

How do I know where actually my data Resides on Data Node?- NICE SOFTWARE SOLUTIONS

What are the major challenges you have faced till now while loading the data into HDFS? - Infosys

Which one is better compression between Lz0/Snappy/gzip? - TCS

What is short circuit in Hadoop?- TCS

What is speculative execution?- TCS

What is distributor cache?- TCS

Explain about high availability in YARN. - HCL

What happens after we submit a request to cluster in yarn architecture? - HCL

During first round interviewer asked me about my Big project experience and current project. Questions around that - Impetus

some coding questions on how to sort the words in a text file using mapreduce and java8 in Hadoop. - Impetus

Have you installed, setup Hadoop in your project?- Impetus

Explain Hadoop architecture and components used in current project. - HCL America

Various types of file formats worked - schema evaluation.- HCL America

Mapreduce core methods, shuffle and sort phase, partitioner.

Size of your cluster, testing, data validation, Hadoop on- prem vs cloud.- HCL America

Scenario based question to identify the storage necessary to handle load of 1TB of daily data with 1000TB of history data - calculate nodes need to be setup for next couple of years - HCL America

Any experience in benchmarking Hadoop cluster.- HCL America

Roles and responsibilities? - Harman

Brief about your cluster?- HCL America

What is your total HDFS capacity?- HCL America

How to enable kerberos?- HCL America

How name node HA works?- HCL America

What is checkpointing?- HCL America

How to set HDFS quota?- HCL America

Day to day activities? - Apple

What is rack awareness?- Apple

How to configure rack awareness?- Apple

Difference between spacequota and setQuota?- Apple

Explain commissioning process?- Apple

How to run HDFS balancer?- Apple

How to set replication factor?- Apple

Started with basic SQL questions like Primary Key, Unique Key, Where clause vs Having?

Explain how mapreduce works, data import/export using sqoop? - Optem

Difference between Hadoop 2 and 3. - TCS

Configuration of your cluster. - Wipro

What are the commands to load and run Spark jar? - Wipro

What is replication factor of Hadoop? - TCS

Maximum replication factor of Hadoop?- TCS

What is the tools used for production cluster? -TCS

?What is the difference between put and copyFromLocal command? - Legato

Yarn architecture explain? - Accenture

Sqoop Interview Questions

If you have the file of 500 GB and 128 is the block size, how many mappers would you launch in your sqoop? - YES BANK

If you have the file of 500 GB and 128 is the block size, how about number of Reducers that you would run in this case? - YES BANK

On what basis will you increase the mappers in sqoop? - Legato

What will you mention --last-value while you are importing for the first time?- Legato

How do you mention the data for incremental last-modified import?- Legato

How do you secure sqoop password? - HSBC

What happens when you increase mappers? - HSBC

What is Eval in sqoop? - TO THE NEW

By default how many mappers and reducers that sqoop runs? - TO THE NEW

What is --direct property in sqoop command? - TCS

What is the difference between --target-dir and --warehouse-dir in sqoop? - City Bank

How to merge two data sets in sqoop? - City Bank

How to fetch updated records in sqoop? - Apple Tech Consultant

Suppose I have 100 tables into the database but I want to import only 5 tables out of 100 tables so how to import only 5 tables using sqoop? - Apple Tech Consultant

How to get unique records in sqoop?- Apple Tech Consultant

I need to import a really large table from Oracle to Hive. How can I design this approach in a way It doesn't fail?

Assume that you are importing data from a 10 TB table in Oracle using 10 mappers and it fails in between after importing 5 TB of data, what will happen now to the partially imported data. How do you handle this? - TCS

You might know that Oracle is case sensitive. For example Table1 and table1 are considered different. Assume that you get a requirement to handle these kind of tables, how would you design a solution using Sqoop, Hive and HDFS?

Let's say that there is an Oracle table and you need to design a solution of daily incremental ingestion to Hive. What factors will you consider for this and how will you design the solution?

Let's say there is a table which does not contain primary key in RDBMS and you need to import that to HDFS. How do you do that in Sqoop?

Let's say there is a table which does contains only non numeric data in a RDBMS and you need to import the to HDFS. How do you do that in Sqoop?

- Accenture

If you have worked on semi-structured data (XML) then how would you perform sqoop import of it. - CTS

How will you import data from RDBMS to hdfs?- CTS

How will you import/Sqoop multiple table from RDBMS?- CTS

If you are exporting data from hdfs to RDBMS and what if the table doesn't exist?- CTS

How the sqoop export work?- CTS

Can you directly import data into Hive internal table.- CTS

How will you import new added data in sqoop.- CTS

How will you import/export sya 101-150 records when you have 200/500 records in table.- CTS

What is the source and destination of Sqoop?

How fast load is done in Sqoop?

Sqoop advantages?

How to connect teradata to Sqoop?

Connecting hdfs to rdbms through Sqoop?

What is the maximum size of data moved?

How to track Sqoop jobs?

What are the parameters Sqoop? - Paypal

How to handle sqoop incremental import automation? - CTS

What is split by option in hive? - Reliance Jio

What are the methods for passing password to your sqoop command. Tell me more secure method other than password file. - IBM

How sqoop all generate data in hdfs. What will be file format by default and what types of files would be there? What if I want import data in Avro format? - IBM

How would you decide number of mappers while running sqoop. What are the parameters to consider while making this decision? - IBM

How sqoop works internally when you run any sqoop import command from terminal.- IBM

Can we import data from RDBMS which don't have a primary key. - Synechron

What is the default database in sqoop? - Cognizant

I have 100 tables, I just want to import 9900 tables. How do I do this in sqoop? -  NICE SOFTWARE SOLUTIONS

Can I use an alias in sqoop query?- NICE SOFTWARE SOLUTIONS

If I do not have a table available in the target system while using sqoop import? Would Sqoop automatically create table while importing data? - NICE SOFTWARE SOLUTIONS

What is sqoop direct method and when to use? - Happiest\_Mind

Suppose you have 10 tables and our requirement is to import 7 tables from there. How would you implement using sqoop? - TCS

Suppose we are doing a sqoop import operation to import 4 tables from Oracle into hive, once import completes in hdfs we see one\_success file and rest are part files. Please tell me the utility of \_success file? - TCS

How would keep your password protected while running sqoop commands?- TCS

How to achieve incremental load with query? - ITC

How data is divided into mappers? - TCS

How many mappers and reducers run in sqoop import by default.- TCS

How to divide the data equally if we have skewed data.- TCS

Write a query to load data from sqoop to directly Hive table.- TCS

Let's suppose we are giving query for loading the data and we don't have any condition or filter the data, then how will we sqoop process the data, will it give any error or will it pass? - TCS

How to import the package in sqoop? -Capgemini

What's incremental append?- CTS

Where does last value an incremental append save, can we view the last value ?- CTS

Sqoop import and incremental append and last modified in where and when used this in your project. - CTS

Difference between sqoop import and jdbc connect? - Capgemini

Why default number of mappers is 4 in sqoop why not 2 or 3? - Alliance Data

Why default number of mappers is 4 in sqoop why not 2 or 3? - Alliance Data

Explain sqoop import / export command? - Harman

Sqoop import gives error "not able to connect to MySQL DB " how to troubleshoot?- Harman

How to increase number of mappers for sqoop job?- Harman

Difference between sqoop import and distcp?- Harman

What is boundary value condition in sqoop? - Altimetric

Explain how mapreduce works, data import / export using sqoop? - Optum

Describe the process of sqoop? - Wipro

How many default mappers and reducers in sqoop? - Wipro

How do you select specific columns while importing data from RDBMS to Hadoop? - Legeto

Suppose you have 80 tables in database, but you want to import only 75 tables. How will you do it? - Legeto

Once you execute a sqoop import, what would happen internally. Explain?- Legeto

Location where sqoop stores its last value.- Legeto

How much data do you receive per sqoop job.- Legeto

On what basis you will increase the mappers in sqoop.- Legeto

What will you mention in --last-value while you are importing for first time. - Legeto

How do you mention data for incremental last modified import.- Legeto

How to delete sqoop job.- Legeto

How is sqoop is used for bringing the updated rows of the table which lies in Oracle? - Bold Technologies

How to decide the number of mappers and reducers in sqoop? - Bold Technologies

**Kafka Interview Questions**

How many messages it can hold in kafka settings? What is size of JMS? - HCL

How many number of  Queues available in your system?- HCL

Pub / sub model in Kafka? - HCL

Disadvantages / limitations of Kafka? - HCL

How can we read data in parallel from kafka? - Sapient

What is bootstrap server? -Vodafone

What is Zookeeper use for?-Vodafone

How do you create a topic?-Vodafone

Do you work on Hive along with kafka?   what is hive ?-Vodafone

What if the one of the replica is crashed / how will you elect leader in this condition? -Vodafone

How you implemented Kafka in your project? can you elaborate? -HCL America

What is the main difference between   kafka  and   flume ?-HCL America

Explain some kafka Streams real- Time Use Cases.-HCL America

Is it possible use kafka without zookeeper?-HCL America

How you implemented stream processing in kafka?-HCL America

How to retrieve multiple column values from kafka topic in Spark structured streaming ? -Iris software

What is consumer group in kafka? - Iris software

What is the role of a leader in kafka ? -Iris software

What is Kafka? - Dassault Systems

Why Do we use Kafka?- Dassault Systems

Explain the architecture of kafka ?- Dassault Systems

What is topic?- Dassault Systems

How does partitioning of topics happen in kafka?- Dassault Systems

How many topics can you create in Kafka? -Boeing

What is consumer record in kafka? - Walmart

How can we maintain security in kafka? -Walmart

Suppose your producer is producing more than your consumer can consume. how will you deal with such situation and what are the preventive measures to stop data loss. -Accenture\*\*\*

Kafka streaming. -TCS

Explain kafka architecture? - IBM

What are consumer groups? - CTS

Why do we need consumer groups? - CTS

How partitions assigned to consumers in a group? - CTS

 What is the flow for the partition assignment?- CTS

What is a broker in kafka? -Legato

Write a code to create a topic?-Legato

How to test is kafka is available?-Legato

How to create a topic? - IBM

Why do we have a redundant set up brokers?- IBM

Write a skeletal code to demonstrate a simple kafka producer to a topic ABCD? - IBM

How do we know if a consumer has consumed a Kafka message or not? - Apple

Can we delete a message from topic once consumer read it? -- Apple

Xxx- Apple

What's your experience in data processing with batch loads and real-time using kafka? Handling incremental and Full Load processes. -HCL America

Explain about Kafka Architecture -  Zookeeper,  Producer, Consumer, Topic, Consumer Group, Pub-Sub model. (Interviewer looking for more on Confluent Kafka) - HCL America

Explain kafka architecture? - Apple

How to create topic?- Apple

What is offset?- Apple

Can we set quota for broker how?- Apple

Explain about  confluent-kafka?- Apple

What are types of API,s in kafka?- Apple

How to debug  the lag issue?- Apple

Cassandra architecture - ring, nodes, CAP Theorem, Consistency Level, Keyspaces and Partitions etc  - Land O Lakes

Cassandra sample sqls -  it does not allow joins. - Land O Lakes

Kafka terminology? - Accenture

Kafka architecture. - Reliance Jio

SQL Interview Questions

2 intermediate level queries based on group by. - Scienapitc Systems

What is indexes in SQL? - TCS

What is the use of dense\_ rank in SQL? - City Bank

Difference between Union and union all? In which approach we prefer to use them? - Lumiq

What is de-normalize table in SQL?- Info Stretch

What is de-normalization 1,2,3?- Info Stretch

What is star Schema in data warehouse?- Info Stretch

Data Lake vs data warehouse- Info Stretch

Difference between Union and Union all?- Fractal Analytics

What is cartesian product?- Fractal Analytics

How to fetch unique elements using SQL?- Fractal Analytics

Difference between where and having?- Fractal Analytics

Lets say I have two tables with only one row in each and same primary key id. If I join them using inner join, what will be the output? - Aligned Automation

Write a query to get the name of manager for each employee. Few questions around the inner join and left join. - Synechron\*\*\*

What is  Coalesce function in SQL? - Walmart

Write a query to find the 3 highest salary in each department in SQL? - Walmart

Complex SQL queries like how to handle updates. - Virtusa

How   columnar file (parquet)  and sequence files get written in HDFS. -Hooner Tekwurks

SQL Query to select the top 3 highest salaries from a Table? - Capgemini

 Difference between inner join and natural join? - Capgemini

How to delete duplicate records from a Table? - Capgemini

General queries like DDL and DML. - TCS

Difference between Union and Union All? - Samsung

Input. -Samsung

Write a program to find the different value between two tables, which have same columns. - Knoldus software\*\*\*

Gave me 2 tables  or data sets of students and marks. Asked me to find students having high score, low score for each subject using sql - Iris software

Asked to solve the above question using Spark additive, data frames and  SQL.- Iris software

Asked me to design singleton class in Java.- Iris software

Tell me about your experience. - HCL America

Standard SQL questions like inner joins, outer join, group by, rank, dense rank. - HCL America

Write a query to find all the batsman score whose current score is greater than previous year score.  - Reliance Jio\*\*\*

Samsung question. - Samsung\*\*\*

Input table ID status  date 100 ordered  8-10- 2020 8.10  100 packet 8- 11- 2020 9. 25   100 shipped 8-13-2920 7.20   100 deliver 8-16-2020 6.05   output table   I'd status valid from valid until    100 ordered 8-10-2020 8.10 8-11-2020 9.24    100 packed 8-11-2020 9.25 8-13-2020 7.19     100 shipped 8-13-2020 7.20 8-16-2020 6.04     100 deliver 8-16-2020 6.05 9-9-9999 9.99     How will you do this?- Samsung

SQL and PL SQL questions. -Land O Lakes

Difference between Delete and Truncate? - Land O Lakes

Different types of views available? - Land O Lakes

Different types of indexes and when each of them is used. - Land O Lakes

Performance tuning. -Land O Lakes

Simple system / DBA tables used. -  Land O Lakes

What is the 3rd  lowest salary in employee table. -Ness

Query to find duplicate of two columns in a table. - NIIT Technologies

About the window function based question asked by giving conditions and using Rank function to solve the query. - NIIT Technologies

Java Interview Questions

Runtime vs Compile time polymorphism, how and where can we make use of them. - City Bank

How to implement abstraction?- City Bank

Diamond problem resolution in Java.- City Bank

What is the difference between Set and map? - Sapient

How to convert array into list?- Sapient

What is the difference between process and thread?- Sapient

Can we implement abstract method in Enum class?- Sapient

What all the different collections in Scala and Java? - Infostretch

How to maintain insertion and read order using map?- Infostretch

What the difference between overloading and overriding? when it is used? - Synechron

What is polymorphism? When is it used?- Synechron

Create an interface for ATM and call it in class.- Synechron

Difference between file Input stream and File reader.- Synechron

Difference between arraylist and linked list?- Synechron

how hashcode() and equals() work? Explain internal functioning of these methods?  - City Bank

Explain Java thread life cycle. - Cloudera

What are the methods available in object class? - Samsung

What is final and finalize in Java?  - Synechron

what is is-a and has-a relationship? How would you implement in JAVA. - JIO

How does JVM manage memory? How garbage collection works?- JIO

Notify() vs  notifyall()  in multithreading.- JIO

Write a program to point below using threads. all three Threads should run in parallel

1

2

3

4

5

6

7

8

9

- JIO

Write a program to reverse a linked list both iteratively and recursively. - Paypal

Write a code to find if given tree is BST.- Paypal

Write a program to find second largest number in an array.- Paypal

Some question from try and final block from exception handling.- Synechron

What is the difference between inheritance and composition?- Synechron

Why we use abstract class when we already have interface?- Synechron

Python Interview Questions

Let's say you have any use case where you need to get the list of files available in a path. How do you approach that in Python? Write a program to find the factorial of a number using tail recursion in Python? -CTS\*\*\*

How to create visualisation charts using Pandas? - Mad Street Den

How to find unique elements in a Python list? - Happiest Minds Technologies

Logic to find how many times a number has been repeated in a list? -  Happiest Minds Technologies

How to get unique elements out of their list? - Aligned Automation

Can you store duplicate values in a dictionary? - Aligned Automation

What is context manager in Python? - CRISIL

What is the difference between tuple and List?- CRISIL

Difference between Call by Value and Call by Reference?- CRISIL

What is list compression, give example?- CRISIL

What is multiple inheritance? Is it possible in java? Is it possible in Python? - CRISIL

How would write a code to find duplicate elements from an array and print them?- CRISIL

What is closure in Python?- CRISIL

How do you package your python code?- CRISIL

How slicing works in Python? - Vodafone

 Explain Python Lambda function. - Vodafone

What is slicing? - Vodafone

Range() vs xrange(). - Vodafone

A non-empty array consisting of N integers is given. -EPAM

How to import csv file using Pandas library? - NICE SOFTWARE SOLUTION

Python and Spark pipelines scenario you have developed. - Land O Lakes

Python Syntax to get word, line count. - Land O Lakes

What is python? - Ness

What is decorator in Python? - Ness

MapReduce Interview Questions

In Map reduce, if we have to find the delta between record and its immediate next record, how will you accomplish that in Map reduce? - TAVANT

Explain the 9 stages of Map reduce? - Apple Tech Consultant

What are the main components of Map reduce job? - Infosys

What are the configurations user need to specify to run the map reduce job?- Infosys

What is Map and Reduce and Chain mapper and Chain reducer?- Infosys

How can we mention multiple mappers and multiple reducers in chain mapper and chain reducer? - Infosys

How many reducers will run when you have 1TB of data with no partition. - Capgemini

What is input split in Hadoop? - Sapient

How number of mapper and reducer tasks gets calculated? - Sapient

Explain map-reduce framework? - Infostretch

Explain functioning of mapper,  reducer, partitioner  and combiner? - City Bank

What is reducers chain? - City Bank

in Map reduce job if you not able to add reducers ,,,what it should do?  How the file should split in MR - CTS\*\*\*

Advantages and disadvantages of mapreduce. - NICE SOFTWARE SOLUTION

Parameters that we pass to mapreduce programs? -Verizon

How the mapper and reducer works?-Verizon

Detailed explanation of mapreduce program. What is the logic in mapper and reducer? -Verizon

Input split in mapreduce?-Verizon

AWS Interview Questions

Which command is used to ingest the data into the AWS s3 service? - Apple Tech Consultant

How to securely give  to a bucket element to a stranger? - EY

How can we copy files between buckets? - Capgemini

What is boto 3? Write down code to least active EC2 instances. - CRISIL

When you spin up cluster in databricks, it gets created at databricks side or AWS side ? - CRISIL

How costing works in databricks and AWS?- CRISIL

AWS Lambda -  what is it? tell me the use case in your project.- CRISIL

Difference between Amazon RDS and DynamoDB. - Virtusa

Which instance you are using in AWS EMR? - Virtusa

Cloud experience in AWS  or Azure (interviewer looking for  AWS) - HCL America

AWS services -  S3, EC2 (types), VPC, Subnet, Key-Value pairs, EMR, Buffer, CloudWatch Metrics - HCL America

Experience in AWS Lambda functions to automate spin up EMR instances, instance type, spot vs on-demand  - HCL America

AWS best practices for storage, security, checking logs.- HCL America

Any experience on working with Docker.- HCL America

AWS S3 bucket security layers. - Optum

**HBase Interview Questions**

Explain the process of insertion in Hbase? - YES BANK

While accessing data from hdfs using CLI we have a  latency while through Hive we have microsecond latency, what happens internally in HBase here?  - Tavant

Row ID duplicates are possible in hbase? - HCL

Assume you are writing data to Hase table.  HBase region server when down. How you will handle this case?. - Capgemini

Assume you have million rows of data in hdfs file. How will you load to Hbase table. - Capgemini

In Hbase, how to copy one table data to another table with different table name - Capgemini

What is Hbase? - Citi Bank

In which scenario / for what type of data Hbase is used? - Citi Bank

Explain functioning of  Hbase. - Citi Bank

 What is column family in Hbase? - Citi Bank

 What is rowkey concept in Hbase? - Citi Bank

 How to upload data in hdfs? - Reliance Jio

 How do you connect Hbase to spark? - Altimetric

Unix Interview Questions

What is the difference between the grep command and AWK command? what is the use of both commands? - Satrix Information Security Pvt. Lmt.

 What is the difference between read, write and Execute permissions? - Apple Tech Consultant

 Assume that you need to proceed with reading a file only when a previous command executes successfully. How you do it? - CTS

 Difference between AWK and sed? - HCL

 How to remove the first ten lines from a file? - HCL

 Write a command to display the files in the directory by file size? - HCL

 Write a command to display the character from 10 to 20 from each line of a file? - HCL

 Streaming file how to search status of processing - Virtusa

 Basic commands - Capgemini

 Command to use to search a particular content in a file? - Legato

Yarn Interview Questions

What are the components in the job flow  of yarn architecture? - Infosys

 suppose you have a very large amount of data may be in exabytes but you don't have the large number of nodes in the cluster then how does yarn handles the resources in such scenario?  - Dassault Systems

 What is yarn and zookeeper? - Citi Bank

 What is client mode and cluster mode yarn? - Citi Bank

 Explain YARN architecture? - Modak Analytics

 How to do YARN tuning? - Modak Analytics

 What happens if node manager fails? - Modak Analytics

 Difference between capacity and fair scheduler? - Modak Analytics

 What is the resource manager you are using? - TCS

Security Interview Questions

How to  create Keytab? - Apple

 Use of kinit and kilist command? - Apple

 Explain kerberos mechanism? - Apple

 What security issues have you faced? - Apple

 How to provide authentication? - Apple

 Security commands? - Apple

 Have you faced with GSS exception how to solve it? - Apple

 Steps to configure TLS / SSL encryption? - Apple

Data Structure Interview Questions

1 Python program -  longest common subsequence- Scienaptic Systems

 How do you find the longest palindromic  substring of a given substring - Power Technologies

 Segregate Even and Odd number in list - ANSR

 Explain internal functioning of hash map. - Synechron

 Write a function in Python that, given an integer N returns the number of integers Which can be formed by the digits of the number which are divisible by X     for example if N = 112 and X = 11 the function should return 2 i.e 11 121     for example if N = 214 and N = 2 the function should return 1 i.e 12,14,24,124,214 - EPAM\*\*\*

 How will you configure the cluster using cloud? - Capgemini

 Write a program to find how many digits are there in a number. - Knoldus software

x

ElasticSearch Interview Questions

First round they ask about my big data experience and current project. - Jio

 questions where around to design  architecture and tools,  Data Structure ETL Pipeline, elastic search on AWS, Kafka, Spark.  - Jio

 second round was about solving coding questions using Java, Spark dataframes like how to create parameter files for metadata of customer data and than sort transaction data using metadata in Spark and Java.- Jio

Cloud Interview Questions

What is Hybrid cloud? - EY

 How to access a bucket through IAM Role? - EY

 If bucket access is there for a user but getting 404 error, what could be the reason? - EY

Kerberos Interview Questions

How to enable kerberos?  - Modak Analytics

 How to create keytab?   - Modak Analytics

 What is kilist and kdestroy?   - Modak Analytics

Snowflake Interview Questions

Snowflake architecture and data storage  - Land O Lakes

How virtual warehouses are defined / created?  - Land O Lakes

ETL pipelines using snowpipe and python - Land O Lakes

Cosmos db Interview Questions

How did you choose partition key in your project - Scienaptic Systems

scaling in Cosmos DB  - ANSR

**SPARK Interview Questions**

Difference between coalesce and repartition can we reduce the partitions with repetition explain which one is best and how - celebel technologies

Difference between map and flatmap? -celebel technologies

When we submit the spark job, when will be the driver program created and where will be the driver program created?- dunnhumby

Below are the cluster configurations(a & b)

a- 1 master node, 5 GB RAM, 10 GB HDD, Dual core processors

b- 2 slave node, 5 GB RAM, 10 GB HDD, Dual core processors

Need to process 30 GB file. Is this cluster configuration is ok? If yes then how many number of executors and the number of cores you will specify for the above configuration while submitting your Spark job? - Accenture

Given the below list how will you create spark sessions? Write the code.

List = [1,2,3,4,5] - Accenture

When you will go for spark cluster mode and spark client mode while submitting a spark job? Explain with a use case and a real time scenario. - Accenture

Programmatically how will you specify the configurations such as the number of Executors and number of cores in your Spark program? - Accenture

While doing Spark submit how you specify the executor memory and no of cpu cores. Please write the full command along the configurations. - Accenture

What is that size of the cluster in your project and what are the memory configurations you are giving while submitting a spark job? - Amid technologies

How you submit your spark code in your project.write the generic command. - Emid technologies

What are the prerequisites for dynamic partitioning.write the property.- Emid technologies

If you want to check your cluster parameters in spark view how will you check that? - Emid technologies

How you move your code to the production environment. - Emid technologies

How you push your code to github. - Emid technologies

Briefly explain your project architecture. - Infosys

What are all the benefits of using spark over mapreduce. - Infosys

How are you scheduling your spark jobs in your project? Which scheduler you used and explain with some scenario from your project? - Infosys

One might be using eclipse or Scala ide while writing the code. You must have got some exceptions in your code. Can you name some exceptions and how you are handling those customised exceptions while performing some operations in spark? - Infosys

You have one multiline json file.how to add your own custom schema to that in spark? - Infosys

Types of dataframe persistence method in explain with a code on how persistence will increase performance. -HCL

How to write data to a partitioned hive table using spark. - IQVIA

Explain lazy evolution in Scala val a=10 val b= 0 lazy val c =a/b print("fdfdfd") print (a) print (b) print (c)At which step it will produce error. - IQVIA

When can we go with sort-merge join or shuffle hash join, can you tell me the conditions? - CTS

How can you join two tables which don't have common joining keys?- CTS

What is your file size in the development environment and production environment? - CTS

What are the production issues you have faced till now?-CTS

What is the difference between repartition(col("a"),10) and df.write.bucketBy(col("a”),8)? Both will work as the same or there is any difference? - CTS

Suppose your Lead was asked you to use spark as ingestion tool instead of Sqoop for getting the data from some transaction table. How do you go about designing this? - CTS

You have 80 tables in your database but you want to import 75 tables. How will you do it? - CTS \*\*\*

How does the state of the last import gates stored for subsequent imports when using incremental import ? - CTS

How distributedby work in hive, given some data tell me how data will be distributed? - Legato

In hive how can you do compression at the field level? - Legato

Since compressed data involves decompressing during processing,how does this increase performance? - Legato

Assume you have a sales table in a company and it has sales entries from a sales man around the globe. How to get top 5 sales based on country and sales volumes? - Legato

If you have only 10 map slots in your cluster,how does a job with 15 mappers run in the cluster? - Legato

In what basis will you increase the mappers in scoop? - Legato

How much of data you deal in daily basis in your current project? - Sistech solutions

How to use DAG, for what purpose we are using DAG, how to use information of DAG as a developer if you have 2000 stages? - fractal analytics

What is the default size of partition in spark?  -fractal analytics

In which cases RDD Will be faster vs dataframe and datasets? - fractal analytics

What is the use of withReplacement? -fractal analytics

Difference between groupByKey and reduceByKey and reduce?-fractal analytics

What transformation in spark you have used which create stages?-fractal analytics

While running spark jobs when you use cluster mode and when you use client mode? - Systech solutions

As you told you were using history data not streaming data in your project then why you used SparkSQL instead of HIVE? - Systech solutions

What is yield keyboard in scala. -Systech solutions

What compression techniques you have used in parquet? -Systech solutions

Do you worked on design patterns? Which design patterns you have used in your project? (Like factory design pattern, Singleton design pattern)? -Systech solutions

Difference between sparkcontext in spark 1 and sparksession in spark 2? - Searce Cosourcing

What were the cluster configurations of your cluster? - Searce Cosourcing

 - Difference between repartition () and coalesce() - Searce Cosourcing

When we use ORC file and when we use Parquet. - Searce Cosourcing

have your job ever failed due to heap memory overhead? how you resolve it? - Searce Cosourcing

what are the optimizing techniques which you have used other than partitioning and bucketing? - Searce Cosourcing

Architecture of spark. -HCL

What is difference between coalesce and partition - HCL

What is RDD? - HCL

How much data you handled in your project? - HCL

Why spark is so much popular? - HCL

What is the difference between data frame API and datasets?- HCL

What is more benefits dataframe API compared to spark SQL? - HCL

Why is spark lazily evaluated? - Yes Bank

Read JSON file flatten it (show the code) - Acceldata.io

Broadcast variable and accumulator variable - NIIT technologies

How to connect spark to redshift ? - NIIT technologies

what is Dist key and sort key in redshift and what is their purpose. - NIIT technologies

What's the difference between spark conf and spark session? - NIIT technologies.

How you orchestrated the spark job ? -NIIT technologies

difference between target directory and warehouse directory in sqoop ? -HSBC

A table transaction with 150k rows is performing badly when the amount is summed by month. How will you go about improving its performance? - HSBC

What is difference between mappartition() and map()? When do you use mappartition() over of map()? - HSBC

How updates will be handled in big data world? Explain possible solutions? - HSBC

what are criteria for choosing different file formats in Hadoop and spark? -HSBC

Can you give me real life examples for lambda architecture? - HSBC

Job scheduling in spark? - scienaptic systems

Garbage collection tuning df = to spark.read.load( filename) how would you optimise this statement to read it faster? - scienaptic systems

Optimisation techniques in spark. - powerupcloud technologies

Explain broadcast and accumulator variables. - powerupcloud technologies

Questions about Delta architecture and lambda architecture. -ANSR

Optimisation techniques in spark. - ANSR

Have you ever used UDFs in pyspark? Can you give an example. Why do we use pandas UDFS--- How do you deploy your Python package to the cluster --- different types of clusters in databricks UI. - ANSR

Let's say you have a csv file of 10th columns, the 10th column is the address column which have a comma separated values. How would you replace the comma separated value with spaces after loading the file into RDD? -Yes Bank

I load 1GB file into memory and then apply the filter operation and then saved the file into HDFS? How many files will be created in hdfs? -Yes Bank

If you are reading from the VSV or text file will you follow the same approach or different in spark? -TO THE NEW

How do you decide the number of executors? How much memory will you allocate to the driver in this case? -TO THE NEW

Let's say that you have an array of elements and you need to remove duplicates from it? How will you do it? Also you need to identify duplicates? -TO THE NEW

if you have a 10gb data table and 100 GB data table which table need to be broad casted? -TO THE NEW

if you have 80 GB of data and 40 GB of RAM how would you run the job in spark? -TO THE NEW

why is parquet the best fit for spark and ORC the best fit for hive ? -TAVANT

how do you populate the values of columns B and C on the basis of column A ? -TAVANT

In what case rdd can be faster than data frames? -TAVANT

in spark DF reader API, is read an action or transformation in spark? -TAVANT

When we have spark SQL then why do we still require Hive? Can spark SQL replace hive completely? -TAVANT

your spark job is throwing out of memory error how do you debug it? -TAVANT

In case one task is taking more time than another, how do you handle? -TAVANT

In a spark job which consists of some transformations and actions, what are all the Java references we expect to see at Drivers and executors? -TAVANT

Let's say you have a Hive table with 100 columns, can you device a de-duplication pipeline if any of the 90% of the columns are matching? Can you accomplish this in Spark? -TAVANT

What is back pressure? -TAVANT

Lets say you want to import data from MySQL, but DBA admin restricted access to only 3 parallel connections at a time.

Suppose you run sqoop job with 10 mappers. What will happen in this case? - Legato

What are the different ways to create a rdd? -IQVIA

What are the differences between rdd dataframes and dataset API? -IQVIA

When you will use dataframe API and when you will go for dataset API? -IQVIA

What happens in the background.  -IQVIA

What is the difference between GroupByKey and ReduceByKey? -IQVIA

What is checkpointing in spark? -IQVIA

How you can drop duplicate records from a dataframe? -IQVIA

Can I have multiple SparkContext or SparkSession objects in a single application? -IQVIA

What is bucketing and partitioning in HIVE? -IQVIA

How you can decide bucketing column on a hive table? -IQVIA

What are the file formats hive supports and which one will give best performance with hive and why? -IQVIA

What are the optimisation techniques you used while working in Hive? -IQVIA

How your are handling incremental data loading in your project? -IQVIA

How to pass password of a database through a file through Sqoop import? -IQVIA

What are the default number of mappers for a sqoop import? -IQVIA

How you can decide number of mappers for a sqoop import to get fast loading? -IQVIA

What are the major differences between SparkContext and SparkSession that you observed? -IQVIA

When we use map and when we use flatmap on spark ? Explain with a use case? - TCS

What is the meaning of paired RDD? - TCS

What is explode function in spark? - TCS

How many shared variables are available in spark? - TCS

How spark gives fault tolerance when an executor will failed?- TCS

How do identity how many jobs/stages will be created for your application?- TCS

What is the use of pipe() operation in spark? - TCS

What are the different types of functions available in hibe? - TCS

What are the window functions you used in your HQL scripts? - TCS

How to make a join as a map side join? - TCS

What are the optimization techniques you used in HQL - TCS

Why ORC is best suited with hive with respective performance? - TCS

What is Hive meta store? - TCS

What is the default delimiter in sqoop? - TCS

How to import all tables excluding two tables from a database in a single scoop import? - TCS

Write sqoop export command for a hive table? - TCS

What is unit in Scala? - TCS

How Scala provide static nature without static keyword like in Java? - TCS

What is the difference between cache and persist? - City Bank

How jobs and stages are divided in a spark application? - City Bank

What is the difference between repartition and coalesce in spark? - City Bank

What are the optimizations we specify in spark-submit command? - City Bank

Explain about spark memory management?- City Bank

What are the configurations needed for spark JDBC connectivity? - City Bank

Explain spark execution flow? - City Bank

What are the optimization techniques you used in your spark code? - City Bank

What are the difference between partitioning and bucketing in hive? How to decide partition column? - City Bank

What are the difference between hive internal and external tables? - City Bank

Write a SQL query to get top 3 salaries in a table? - City Bank

Write a sqoop import command and explain the parameters. - City Bank

What is case class in scala? - City Bank

What is difference between case and object? - City Bank

Write and explain the program for palindrome program in Scala? - City Bank

Explain Scala execution mechanism? - City Bank

Explain the difference between list and tuple in Python. Project scenarios to explain when yo use tuple and list.- WALMART LABS INDIA

Write a word count program in word and sort the words in ascending orders.- Accenture

I have a requirement to do set of spark ETL operations. At which circumstances will u use dataframe and when will you use dataset? - WALMART LABS INDIA

What all optimizations we can do while extracting data from SQL server/Oracle using spark. - WALMART LABS INDIA

How many partitions will be there if I use spark.read.jdbc to read from an oracle table. - WALMART LABS INDIA

Catalyst Optimizer's role.- WALMART LABS INDIA

Which face will cost based optimization happen in spark. - WALMART LABS INDIA

I have a parquet file named employee in HDFS path.file is having employee information with manager id.spark script will read Parquet file and do a self join to get the manager name.how many partitions will be there in the dataframe which came after the join.what all factors will determine the partitions in this scenario. - WALMART LABS INDIA

Sql Question 1

Ticket#,Task#,Status

1,A1,Completed

1,B1,Completed

1,C1,Inprogress

2,A1,Completed

2,B1,Completed

2,C1,Completed

3,A1,Completed

3,B1,Inprogress

3,C1,Inprogress

Find out all ticket numbers having only A1 task as completed. In case it should print out Ticket#3.

Do this in spark dataframe operations and sql. - WALMART LABS INDIA

Sql Q2

emp sal mgr

1, 2000, 10

2, 5500, 10

3, 1000, 20

4, 1500, 20

7, 20000, Null

10, 5000, 7

20, 2000, 7

Find out managers who are getting salary above three reportees (both spark dataframe and sql) - WALMART LABS INDIA

Speculative execution in spark. - WALMART LABS INDIA

Read a text file in spark rdd to print the number of times spark is present there. How the stages will be created and decided by spark?

How the Tungsten engine helps to speed up the spark execution?

The internals of catalyst optimizer?

How to implement bucketing in Hive, decide the number of buckets, and how it helps in the Hive in execution?

Questions on my project, its component, parallel processing in spark?

Sqoop direct method and when to use?

How to access different versions of S3 bucket in spark? - Mindtree

What is water-mark in spark structure streaming? explain it with an real- time example. - Mindtree

How the tungsten engine helps to speed up the spark execution. - Hppiest\_Mind

The internal of catalyst optimizer and how the query lifecycle has been managed by catalyst optimizer. - Hppiest\_Mind

1: Refer the below table and arrange the records as per the max(sum) of f\_count in city wise.

2: How will you have the total sum and cumulative sum of f\_count for each city, how the output will display in both the cases. - Sapient

c\_name f\_name

hyd 10

Mum 20

Ban 12

hyd 34

mum 23

Del 22

Hyd 24

Ban 28

Create a dataset having name Salary dept. Display the max 3 salary for each department using Spark core and rdd only (no dataframe and SQL will be used) - Sapient

Create a dataset having name Salary dept. Display the max 3 salary for each department using spark core and rdd only(no dataframe and SQL will be used)- Hppiest\_ Mind

What is Trigger means in spark structured streaming and will 5 seconds trigger means spark will not receive the data for 5 seconds?  - Hppiest\_Mind

How the partitioning work in spark structured streaming with Kafka. - Hppiest\_Mind

When we run a spark-submit job then on which factor we decide the number of executors? Suppose I want to process 100GB of data then how many no.of executors are required to process it? Which logic is used for it? - Satrix informations security Pvt.Lmt

Which logic use in spark to define no of stages? - Satrix informations security Pvt.Lmt

Why RDD called a lazy Evolution in nature i.e what is a particular reason for it? - Satrix information security Pvt.Lmt

Explain the lazy Evolution in spark? - Satrix informations security Pvt.Lmt

Suppose I load one file into RDD & before performing any action I delete that file from the background.so whether it will show any error or run successfully without any error? - Satrix information security Pvt. Lmt

What is the difference between flatmap and mappartition? - Satrix informations security Pvt.Lmt

Suppose one table in the hive has 100 partitions & when we connect to spark and load the same table into spark then how no. of the partition will get generated in Spark? - Satrix information security Pvt. Lmt

Suppose you are writing one data frame in the HDFS file name as s1.csv & in the destination s1.csv file already exists. So how will you ensure that will not give any compile-time & run time error? So what precaution you will take to write the data to the destination? - Satrix information security Pvt.Lmt

Input data. -Lumiq

Employee table. - Lumiq

When we submit spark jobs so how it will works internally? - Lumiq

What is the difference between narrow and wide transformation? - Lumiq

What is the difference between spark context and spark session? - Lumiq

Suppose I read a CSV file in a spark that has id & name column but the file has some extra commas in data so how do deal with extra commas?  - Lumiq

The architecture of spark in detail? what is driver, stage & task in spark? - Apple Tech Consultant

What is the difference between GroupByKey and ReduceByKey? - Apple Tech Consultant

How to add the column into existing dataframe? -Apple Tech Consultant

What is the main difference between HBase & Hive?

Explain partitioning in Hive?

Explain bucketing?

Difference in partitioning and bucketing?

When to use partitioning ?

Sqoop eval command usage?

Explain bound val query.

What is split by in Sqoop?

Anonymous function in Scala, explain?

What is partially applied function?

Why Spark not MapReduce? What is YARN?

What is the language supported by Spark?

How do you use Spark Stream?

Use of cat command, grep and sed?

OLTP vs OLAP?

Explain the ETL process?

Analytical functions usage? - CTS

What are transformations and Actions in spark. - City Bank

Dataframes vs Datasets. - City Bank

Why spark is faster as compared to MR? - City Bank

Reduce vs ReduceByKey.which is better and why? - City Bank

What are the important parameters in spark submit command. -Infosys

What are the optimization techniques used for sqoop in your project. - Accenture

What is the development environment file size in your project? - Accenture

What is tail recursion in scala? - Capgemini

How do you access tables in hive from spark? - Capgemini

What is the difference between RDD and dataframe & dataset? - Capgemini

What is the difference b/w Izo and gzip compression techniques and which compression you used in your project and why? - Capgemini

How you will take the moving average of stock data with a window size of 3 rows. - Hopiest\_Mind

What is Hive bloom filter and implement with an example. - Mindtree

Explain globalView/tempview. - Imagine Tech Services

Which file format you have used? - Imagine Tech Services

Why Parquet - Imagine Tech Services

How you submit the job? - Imagine Tech Services

Explain executer and cores. - Imagine Tech Services

Lets say that there are 10 executors and each with 6 cores each. How many tasks can you run in parallel?

Assume that the file is in HDFS and its size is 100 GB and you are loading that file to a dataframe. How many partitions will the dataframe contain and how do you check that?

Assume you have been given a use case of importing data from a table in RDBMS and find the distinct count of values in a column. What tools you will use of this use case?

Assume that you are working in an environment of Python and Spark. One fine day, your client/PM hears about dataset in Spark that it can detect error in compile time itself, etc and insists you to use dataset in Pyspark. What would be your response?

Assume that your workflow is Spark is too long and they are chained up till an action is executed i.e. comprising of more than 30 transformations and a final action. What feature on Spark you can use here to avoid recomputing the whole flow on occurrence of any task failure in a node?  - TCS

I have a text formatted file in HDFS. I need to move this file to a Parquet table. How can I do that?

What is the difference between Avro, Parquet and ORC formatted files? Give me some scenario where you can go with each of these types?

Which format do you recommend for a high read intensive and high write intensive operations? - TCS

Explain the overall in- memory process after we submit spark job. - IBM

Explain stages and tasks. - IBM

What are narrow and wide transformation. - IBM

What all production issues you have faced and how you resolved? - IBM

What optimization techniques you have used? - IBM

In what all condition we need to change driver memory. - IBM

Which join you will do when the data is skewed? - IBM

Assume that you are working in an environment of Python and Spark. One fine day, your client/PM hears about dataset in Spark that it can detect error in compile time itself, etc and insists you to use dataset in Pyspark. What would be your response?

Assume that your workflow in Spark is too long and they are chained up till an action is executed i.e. comprising of more than 30 transformations and a final action. What feature on Spark you can use here to avoid recomputing the whole flow on occurrence of any task failure in a node?

How do you handle skewness of data in Spark?

What determines stage boundaries in spark? - CTS

Lets say you have created a Spark application but you are unable to read the tables from Hive metastore? What should you check here?

Can you do dynamic partition using Spark like the way you do in Hive?

What do you think are the reasons why we see SparkSession as entry point to your Spark application in Spark2 and not SparkContext as in Spark1?

Assume I have two tables TableA and TableB both being table of 1 TB each and the joined output need to be written to a Hive table. But job is failing due to memory issues in the cluster when you join these tables together. What do you think could be done to optimize this join? - Accenture

How you do data cleansing in your project? - Alliance Data

What all transformation you have used in your project? - Alliance Data

What is broadcast joining? - Alliance Data

We want to add a column which is the sum of average in bonus salary.how we can achieve this? - Alliance Data

How you do join at spark dataframe level? - Alliance Data

How to build partition in hive table on spark ? - Alliance Data

How to load hive table in spark? - Alliance Data

What is the direct method in sqoop and where is implement ? - Mindtree

How to import multiple rdbms tables to multiple hive location in sqoop? - Mindtree

What is pair RDD? Especially in Joins, where we use them? - Tavant Technologies

How to do checkpointing in spark? - Tavant Technologies

Is logging is supported in spark standalone mode? How?[what is that framework?] - Tavant Technologies

Spark program has 3 input file, data in these 3 files are numbers from 1 to 1 billion number, i.e. (1,3,4,2,6,7,8,9,5…. till 1 billion)

Now write missing numbers logic in optimized way in any logic [Missing numbers should be saved in DataFrame].  - Tavant Technologies

There is a legacy MapReduce program, now there is an enhancement to do, but instead of MapReduce, you need to do it in Spark, But you shouldn't rewrite MapReduce code to Spark code, in this process, can you write your enhancement code in Spark & make it invocable in MapReduce code? If yes, how is it possible? & how does it work?  - Tavant Technologies

Spark core: We have 200 cores, 6GB RAM memory, now if we have launched spark program, how much memory will be used for 300 core executors for submitting job? -Tavant Technologies

Cluster has 200 cores, 1000 part files as input 16 GB RAM & how many executor cores are required? - Tavant Technologies

In another scenario, 100 part files, 200 cores, 16GB RAM & how many executor cores? - Tavant Technologies

What kind of different API's,You have used till now with spark? - Tavant Technologies

Difference between RDD vs dataframe and dataset. -HCL

Types of variables in spark.- HCL

Hive optimization techniques used in spark?- HCL

What are spark versions used? -HCL

How to write spark SQL program using following sales data: customer\_id, customer\_name, sale\_amount, date.

Output: customer\_id, namem, total\_sales\_in\_current\_year, total\_sales\_ib\_current\_quarter, total\_sales\_in\_prev\_year, total\_sales\_in\_prev\_quarter

Dataset:

customer\_id, customer\_name, sale\_amount, date.

1001,peter,100.28,2018-10-20

1002,parker,200.35,2018-10-21

1003,ravi,512.00,2018-11-21

1001,parker,300.05,2018-10-21

1009,mark,99.99,2018-11-10

1010,suraz,0.99,2017-09-25

- Capgemini

I have two datasets, one with 100 million records and one with 20 million records. We would need to have a join between these two datasets and we would need to select tiny executors or fat executors. Which one is the best for this scenario and why? - Capgemini

What is the difference between dataframe dsl and SQLquery? - Capgemini

What is the generic framework in spark? - Clover Infotech

How do you debug failed spark jobs? What do you do incase of spark job failure? Will you return it? How did you get notification about the same? - Clover Infotech

How do you managing your memory configuration in spark? How do you decide how much memory and how many cores needs to assign to specific spark-submit program.

How these specific parameters are used like what is this executor memory and executor cores - Clover Infotech

If you want to caching between spark jobs how do you do that? - Clover Infotech

You have 100GB of table and 1 table of small table.how did you join them in spark? - Clover Infotech

What is windowing functions and skewed joins in hive. -Trunkoz.com

How do you increase the mappers dynamically in hive. -Trunkoz.com

How did you define on which column partitioning need to be done if customer didn't specify any column? - Trunkoz.com

Once the data is imported using sqoop from RDBMS, how to test it whether all the data got imported or not?

How to do column to column checking. Very Important question. -Trunkoz.com

How to do incremental load into hive when there is no stamp and date in the source? -Trunkoz.com

If suppose I have 100 rows in 1 RDD. -Trunkoz.com

Why you get out of memory exception in spark? Explain in detail java.lang.OutOfMemoryError: GC overhead limit exceeded and/or a java out of heap exception. - IBM

Diff b/w group by key and reduce by key in spark. - IBM

How to find max of sal in SQL.-IBM

How do you find the average of given list in scala ? -IBM

How do you updates the records in hive? -IBM

What formats you worked on?In Parquet how do you update a record. - IBM

Cost based operation is happening on dataframes and not on datasets, why? - IBM

If input file comes with diff delimiters every times for few columns in a table, what file format you use?

How to map Hive table with Hbase table to map data

hive> CREATE EXTERNAL TABLE hbase\_table\_2(key int, name string)

> STORED BY ‘org.apache.hadoop.hive.HBaseStorageHandler’

> WITH SERDEPROPERTIES (“hbase.columns.mapping” = :key,cf:name”)

> TBLPROPERTIES(“hbase.table.net” = “shakti”)

 - IQIVA

Write word count program in scala. - IQIVA

How to map hive table with Hbase table to map data

hive> CREATE EXTERNAL TABLE hbase\_table\_2(key  int, name string)

>STORED BY 'org.apache.hadoop.hive.hbase.HbaseStoregeHandler'

>WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,cf:name")

>TBLPROPERTIES("hbase.table.name" = "shakti") - IQIVA

How tune spark job? -IQIVA

How did distributed computation works before Hadoop?- Sapient

Write spark streaming code to read a data from kafka and write it to hive. -Sapient

How to write very big file in Cassandra using spark? -Sapient

How right is being happen in Cassandra? -Sapient

Suppose you have one service which is using particular port number, and we have to assign that port number to another service. How do you change the port number? Without disabling anything, we have to make it. - Cloudera

In Linux, if I have two hosts in one network, if default gateway is enabled in one host and not in other, will it able to send messages from one host  to another? - Cloudera

I am trying to copy some file from one location to another, but could not find some files, how can I find it? -Cloudera

Quick introduction -Vupico

Can you please tell me which tools you have been using for data ingestion pipeline apart from sqoop? -Vupico

How are you pulling the data from S3? -Vupico

Using scoop what is your destination like dump it in HDFS and then use spark on it or dump it from Hive tables? - Vupico

In all this ecosystem, which tools you have been using and relatively exploring in? - Vupico

Cloudera provides which implementation which is similar to HIVE? How it is different? -Vupico

Have you worked with incremental loads in scoop and How do you handle basically import and export? - Vupico

Let's say while doing sqoop, my job is not writing directly into hive but it is loading the data into HDFS and somehow I am trying to fetch that data I am getting a date along with the timestamp. But my requirement is that in Hive I need to have the partition key based on date not based on timestamp. So how will you handle this kind of scenario? -Vupico

I have a Scala class which is doing some RDD based operations and at the end in terms of action it is basically dumping the output into HDFS location. I compiled and create a jar and then I do Spark submit. So what all things happening in the background? -Vupico

How the stages will be determined? How the number of tasks will be determined and so on? -Vupico

In the past experience have you ever come across the term Uber mode also while checking logs at the other end you may have come across the term? So what is Uber mode? -Vupico

I am trying to read the sequence file and then basically trying to execute that using Spark sequel and dumping the data either into Hive or into some other external location. Another user implemented the spark SQL file and they are using spark-sql command for executing. If you are the architect in that place which option will you approach to do? -Vupico

What was your role in handling Kafka in your project, does your project use Kafka tool for ingestion? What did you achieve in this role? -Vodafone

What is Kafka and explain its architecture? -Vodafone

How do you check the status of consumer?-Vodafone

Tell full procedures and commands of creating topics, producers, consumers?-Vodafone

What is the difference between Groupby Key and reduceby Key? -CTS

Why cache as spark already provides inmemory feature? - CTS

What are the different optimization techniques in spark? - CTS

What is command of spark submit? - Paypal

What is a smart version you worked upon? -Paypal

5 transformations and actions you know in spark. -Paypal

What are the analytics functions in spark.-Paypal

Explain spark Architecture?

What is RDD?

What is the role of coalesce () and repartition () in Map Reduce?

How do you specify the number of partitions while creating an RDD? What are the functions?

What are actions and transformations? What are the problems that you have faced with Spark and how did you troubleshoot them?

Components of spark ecosystem?

Performance tuning in spark? - American Express

What is the difference between driver memory and executor memory? - Sigma Systems

What is the fraction of executor memory divided to store any input file and do the computation? - Sigma Systems

Why do you use dataframe instead of RDD?- Sigma Systems

What is the difference between saveAsTable() and insertInto()? - Datamatica

How json file infer schema while loading as dataframe? -Datamatica

If our highly optimized spark query runs but still failing due to out of memory exception then what need to do? -Datamatica

How to read Data from a CSV file with and without header. - Legato

Did you used Hive? Where do you used? - CTS

What is Kafka consumer and producer? Explain?- CTS

Can you explain YARN architecture and how it works? -CTS

What do you mean by commodity hardware? - CTS

What are the tools used to connect with edge node in Hadoop? - CTS

Did you worked any integration with Kafka ?  -CTS

What is the need of data locality in Hadoop? -Capgemini

Did you worked with SparkSQL and what are the differences between SparkSQL and DSL ? -Capgemini

How you implemented spark streaming ? -Capgemini

What is WithColumn in DSL and explain how it used ? -Capgemini

Did you worked on any  integration with spark streaming explain? -Capgemini

How to save my dataframe to the hive table with partitions? -CTS

How to see the average highway speed of cars by class and years? -CTS

Which is the default join in spark. - Legato

Tell me the Syntax for broadcast join. - Legato

Which is the default join in spark. - Legato

Which spark achieves parallelism? - Mad Street Den

What is tungsten format? -Mad Street Den

Spark optimisation techniques ? -CTS

difference between repartition and coalesce ? -CTS

Why spark rdd is immutable ? -CTS

How to extract data from RDBMS in spark ? -CTS

What are the security options in spark? - Mad Street Den

What are the transformations and actions you have performed in your project? - Mad Street Den

What kind of tool you used for automation of spark jobs? Explain that?- Mad Street Den

What is DAG? - Mad Street Den

What is UDF? How does it work? - Mad Street Den

How to create new columns? - Mad Street Den

Why/how RDD are resilient? - Mad Street Den

What is your cluster size? - Mad Street Den

How to run CLI commands from inside Pyspark script? - Mad Street Den

Spark JDBC connectivity details - Walmart

Why stages are formed in spark? - Walmart

Difference between reducebykey and groupbykey.- Walmart

What is garbage collector in spark? how you assign memory for it. - Walmart

How to overwrite the file in spark in same location where is the file is read? - HCL

What happens if the data I am reading is larger than my memory? - EY

Difference between Avro and Parquet? - EY

How do you deploy a job in EMR? - EY

How to increase executer/driver memory when submitting a job? - EY

What is columnar data storage? - EY

What is window function?- EY

Difference between window and groupby function ? -EY

Let's say new data is continuously pushed to S3, How can you take and move to another s3 as it comes? - EY

Can we store data in EMR's HDFS? if yes why not use that? - EY

Explain your project? - EY

What does Attunlty replicate do? - EY

How RDD partitioning works? -fractal analytics

Difference between Dataframe and Dataset? -fractal analytics

What are the different ways to optimize/increase spark performance? -fractal analytics

Difference between client and cluster deployment mode? -fractal analytics

Internal components of YARN? -fractal analytics

How does YARN work with Spark? -fractal analytics

If dataframes df1 and df2 are there, How do you take common rows as well as all the rows from df1?  - fractal analytics

What is your typical day to day work? -fractal analytics

What does ReduceByKey do? -fractal analytics

What is flatmap function? -fractal analytics

How to find the count of unique words from corpus of texts? -fractal analytics

How to configure compression with spark? -fractal analytics

how to pick a particular row from a spark data frame ? - Happiest Minds Technologies

How to transform corpus of texts to single words per line?  - Happiest Minds Technologies

Difference between filter and where function ?  - Happiest Minds Technologies

What is an overhead ?  - Happiest Minds Technologies

How garbage collection works ?  - Happiest Minds Technologies

Can we optimise/increase performance through serialization ?  - Happiest Minds Technologies

How bucketing works ?  - Happiest Minds Technologies

Different types of pyspark joins ?  - Happiest Minds Technologies

where an accumulator can be used ?  - Happiest Minds Technologies

Is there an optimization technique available for parquet or other types?  - Happiest Minds Technologies

Can we use HDFS to store while using EMR ? - Aligned Automation

how to join two data frames ?- Aligned Automation

How to create new columns with a value of one column. [Value is present like A,B,C,D] - Aligned Automation

How spark is quicker than mapreduce? Is not writing to the target right away after computation is the only reason?  - Aligned Automation

Difference between coalesce and repartition ? Which one is an expensive operation? - Aligned Automation

How to avoid shuffles ? - Aligned Automation

Difference between broadcast and shuffle-merge join? - Aligned Automation

Difference between cache and persists? - Aligned Automation

What are Broadcast variables ? - Aligned Automation

What could be a reason for spark out of memory error ? - Aligned Automation

Does mesos have any advantage over yarn? - Aligned Automation

Explain your project? - Capgemini

what are broadcast variables? -Capgemini

Does accumulators are used anywhere else other than to keep a synchronised count? - Capgemini

What is lineage graph? -Capgemini

Is data from an RDD? -Capgemini

How spark dataframe differs from pandas dataframe?-Capgemini

Where driver manager resides? -Capgemini

Does location of driver manager matter in spark performance?l-Capgemini

What does rank function do?-Capgemini

What does row\_number function do ? - EY

Have you done anything to improve your projects performance?-Capgemini

What is the default join in spark? - Legato

What is the default delimiter in hive?- Legato

Write the syntax for joining two DF using broadcast?- Legato

How you will convert DF into RDD?- Legato

How do you decided the number of cores versus the number of executors? - Happiest Minds Technologies

What is data skewness in spark and how to overcome that. - Happiest Minds Technologies

Difference between Hive static and dynamic partition?-Happiest Minds Technologies

Explain partition in kafka? -Happiest Minds Technologies

Why spark is preferred over mapreduce?Happiest Minds Technologies

How to do an incremental load into Hive table?- Happiest Minds Technologies

How you can calculate the executor memory? - Happiest Minds Technologies

Which is the best file format that you can use for Hive tables? -Happiest Minds Technologies

How you are dealing with updated in your data lake? -Marlabs

How you are handling connections strings in spark when reading data from blob storage. -Marlabs

Explain different file formats and advantages?-Marlabs

Internal difference between spark session and spark context?-Marlabs

Explain some challenging scenarios that you have faced in your project?-Marlabs

The main difference between dataset vs data frame ?-Marlabs

How to resume if the spark streaming job should fail? - CTS

What is unit type in Scala?- TO THE NEW

What is option in Scala and how to use that in your application ?- TO THE NEW

What is functional programming in Scala and how it is different from oop? what do you mean by pure function in Scala?  - TO THE NEW

What is match in scalar and write a code for this?- TO THE NEW

Partitioner in spark and how to create a custom partitioner in spark and what is the benefit of it? - TO THE NEW

Foreach is an action or transformation in spark, you need to justify your answer if you say It is an action or transformation? - TO THE NEW

Spark rdd operation only to find out the max salary for each department?- Iris Software

Difference between rdd caching and dataframe caching?- Iris Software

Difference between spark and mapreduce?- Synechron

Why is spark preferred?- Synechron

Describe spark architecture and functionality?- Synechron

What are the names of the two files that are used by namenode to store filesystem information? - Synechron

Why and how replication factor is used?- Synechron

Command to copy data into hdfs from local.- Synechron

How would you write a classification program in spark using Java?-Dassault Systems

Difference between rdd, dataframe and dataset?-Dassault Systems

How to remove null value from an rdd?-Dassault Systems

Write a spark program to select few columns, and then filter out few records?-Dassault Systems

Give an example of data skewness that you have come across while developing a spark application and explain how you deserve it. -Dassault Systems

How would you optimise a spark job?-Dassault Systems

Difference between repartition and coalesce?- City Bank

What settings are required for setting up a spark cluster and executor? Where we do these settings? - City Bank

What is executor? - City Bank

How would you copy data from spark to hive?- City Bank

How to execute and submit spark job?- City Bank

In spark SQL how to save new data only in append mode?- CTS

The difference between rownum and dense rank?- CTS

What is partial ordering in scala and explain with an example?- CTS

What is raise ValueError in spark and explain with an example?- CTS

What is function currying in Scala and explain with an example?- CTS

Write the spark code to access S3 bucket?- CTS

What are the coding optimisation you did in your spark code?- CTS

What is tree reduce and explain with an example?- CTS

How to add a new row to a dataframe (ans:use flatmap to add row)?- CTS

how to read 1 tb data in 256 GB RAM?- CTS

How to read multiple topics in spark structured streaming with kafka?- CTS

What is cluster configuration in your project?- Lumiq.ai

What is S3 consistency?- Lumiq.ai

In athena do we create internal table or external table?- Lumiq.ai

Difference between intersect and inner join?-Dunnhumby

Explain spark architecture from end-to-end?

What happens when file size is more than file in Spark? In this case what we need to do?

What are the difference performance techniques you have followed in spark?

- Global logic

What is higher order function write code for it?- Global logic

How to find the missing number in the following array?- Global logic

What is the spark cluster used by aww glue?- Global logic

Scenario based question?- Clairvoyant

What is the difference between hashmap and hashtable difference?-Clairvoyant

We need to convert given string into key value pair like below i mentioned

input: -aa#&kls\*bb&^%lkj@cc%#asd

output: -aa->kls

-bb->lkj

-cc->asd

 -Clairvoyant

input:

id Name department salary

2 AA Science 10000

5 BB Maths 20000

8 CC Science 14000

10 DD Bio 13000

output:

maximum salary by department. -Clairvoyant

Please explain spark internals?- fractal analytics

No users were creating files in hdfs but automatically some hdfs files are creating, how will you debug?  - Cloudera

ETL process for data optimisation in detail - WALMART LABS INDIA

Intra load balancer?- TCS

Spark query is running slow how you will debug?- Cloudera

Yesterday one query in Hive ran successful and the same query today it is taking much time and what are all the reason you may find? - Cloudera

Tungsten in spark?- Cloudera

You have two servers each 50 GB RAM and for one yarn job container size needed is 70GB how internally the container sizing happen. - Cloudera

Life cycle of yarn job ? - Cloudera

Which Hadoop distribution work upon? - WALMART LABS INDIA

Hdfs CP giving some copy error what can be the causes? - WALMART LABS INDIA

What are the data formats that hive store?-WALMART LABS INDIA

Explain the concept of high availability in Hadoop. - WALMART LABS INDIA

How would you allocate more memory to the application container in yarn? - WALMART LABS INDIA

What are the side data distribution techniques? -WALMART LABS INDIA

What if job tracker machine is down? -WALMART LABS INDIA

Where are Hadoop's configuration files located? -WALMART LABS INDIA

What is the difference between input split and hdfs block? -WALMART LABS INDIA

What is jmx?- TCS

Hive metadata where it stores ? -TCS

How Hbase write works? - TCS

What is data localization? - TCS

Difference between zk quorum and ensemble?- TCS

You are writing in hdfs and suddenly failover happens what will happen in the backend? - TCS

Scenario: Let's say for e-commerce application, I will be getting files in S3 for every 15 min and I need to run the job. How do you build the batch pipeline and what tools you use? - Boeing

 Given one Kafka program and asked to persist the data in MySQL using spark streaming? - Boeing

How will you handle security in AWS cloud?- Boeing

Why you choose DataFrames instead of Datasets? Since datasets provide more type safety than dataframe? - Boeing

AWS EMR has both HDFS and S3. When you will choose HDFS over S3? Where would you store all your temporary data while executing jobs in cloud? - Boeing

We are getting sensor data from IOT devices. we have to build Streaming pipeline to process IOT sensor data. How do you go about architecting the pipeline for streaming? Discuss tools you use? - Boeing

Difference between spark streaming and structured streaming.- Reliance Jio

 Difference between partitioning and bucketing in Hive, when to use partitioning or bucketing. - Celebel technologies

 Suppose if we give number-of-buckets as 5 and partition column has 4 values, then how many files will be created? - Celebel technologies

Explain spark architecture.- Celebel technologies

What is broadcasting in spark?- Celebel technologies

Suppose we have 10 GB file and we are having 4 nodes and 100 GB memory. now if use cache and broadcasting on the 10GB file. How much memory it takes in both cases? - Celebel technologies

What are the types of transformation in spark? Explain them. - Celebel technologies

What is the difference between internal and external metadata? - Celebel technologies

Suppose if we read a file, will it is a lazy evaluation in spark? if yes, then in dataframe how it will store the metadata? - Celebel technologies

Difference between Union and union all. - Dunnhumby

 What is intersect and union?- Dunnhumby

What happens when join is done in SQL?- Dunnhumby

Distinct one irrespective of source and destination

Input:-

source destination

hyd,del

del,hyd

ban,hyd

hyd,del

output:-

ban,hyd

- Dunnhumby

What is lazy evaluation and what is the importance of it?- Dunnhumby

What is data skewness, how it will affect the job. - Dunnhumby

Difference between csv, parquet, orc, avro in spark. When you prefer parquet and orc? - Dunnhumby

Scenario based question:-

how csv file will be read. for example we have 3tb of memory and 1tb of csv file. How it will be read? - Dunnhumby

Optimisation technique in Hive and optimisation technique you have implemented in Hive? - CTS

I have 100 records and I want to export these 100 records in 2 transactions, how will you do it in hive and sqoop? - CTS

Location where sqoop stores its last value?- CTS

Per scoop job how much data do you receive? - CTS

On what base tasks are created in DAG? - fractal analytics

Can we use bucketing without partitioning ?- Clairvoyant

What is the significance of number of buckets which we mention in hive? - Clairvoyant

We have created Hive external and internal table with the same location. Now if we delete the hive internal table what happens? - Clairvoyant

What is spark partition? - Clairvoyant

What are the different types of file-formats explain them? - Clairvoyant

What are the different types of modes in spark, and what is the difference? - Clairvoyant

Have you worked on multi threading in Scala? - Legato

What will you mention --last-value while you are importing for the first time? - Legato

How do you mention the date for incremental last modified import? - Legato

Let's say you have created a partition for Hyderabad but you loaded Chennai data, what are the validation we have to do in this case to make sure that there won't be any errors? - Legato

Consider I have a Hive table with 100 column data and it is an historical data then how will you device a de-duplication pipeline and that de-duplication pipeline will be considered only when we have any 90% overlapping columns (any 90 columns out of 100 columns) between the records and then only 1 record will filter out? - Systech Solutions

I am getting memory out of exception using groupByKey and want to resolve it but do not want to apply any other transformation. How would you do that? - TO THE NEW

Have you done any optimisation in spark? - Cognizant

Difference between map and flatmap ? which is beneficial? -Cognizant

How to find difference between data frames?- Cognizant

What is broadcast variable and accumulator?- Cognizant

How to create an external table from spark dataframe?- IBM

What is the significance of num-executors, when we give 10 executors would Spark force Cluster to run 10 executors always? What is the thing that will decide how many executors should be run at that time? - IBM

I have dataframe A and dataframe B. Give me data frame which contains records present in A but not in B. - IBM

How exactly spark job runs on a cluster? What is different in yarn mode and and standalone mode? - IBM

How do you calculate the cumulative sum in spark? - IBM

How to save spark dataframe as Hive table? - IBM

Please explain whether saveAsTable() creates external table or internal table. - IBM

Please write down code to create your own custom schema for dataframe.- IBM

How to decide the number of executors in a cluster? -Walmart

How to add new columns in DF? -Walmart

Explain encoder in spark. -Walmart

Difference between dataset and dataframe? -Walmart

Read emp.csv and manager.csv into dataframe and give me dataframe which has only managers who are employee as well. - Walmart

EMP.CSV

EMP ID Manager ID

1 7

2 1

3 2

4 6

Manager.CSV

Manager ID Manager NAME

1 Shubham

1 Shivang

1 Sakshi

1 Shruti

1 Saksham

1 Manik

1 Mukesh

What is speculative execution in spark? - CRISIL

How join operation is performed in spark ? Take two dataset and write down code to perform join operation. - CRISIL

What is accumulator? How does it work?- CRISIL

How catalyst optimizer works under the Hood?- CRISIL

How join operation works under the Hood?- CRISIL

If your code is using DataFrame API and using spark SQL, is there any performance difference? - CRISIL

What is merge sort join and hash join? - CRISIL

How would you read JSON file into dataframe?- CRISIL

What is streamingcontext explain in detail? - WALMART

What is block interval and batch interval in spark streaming?- WALMART

How to achieve rebalance in kafka? -Wipro

rdd1 -> (filter)-> rdd2 -> (map) -> rdd3. Now I deleted rdd1 from cluster, now if I run action on rdd3, Would it work? - Vodafone

 What is the difference between groupbykey and combineByKey? - Vodafone

 What is sparse matrix in spark?- Vodafone

 I have 200 GB file, I am reading that into spark cluster and I am getting out of memory error. How would you solve this issue? - Vodafone

What is difference between repartition() and coalesce()- Vodafone

How to write a file into parquet format using dataframe in spark?- Vodafone

I have 3 nodes and 15 GB file, How would I partition data equally in these 3 nodes in spark cluster? - Vodafone

How do you deploy your Spark-Scala code in production cluster? - Vodafone

Difference between wire transformation vs narrow transformation - Vodafone

RDD API vs Dataframe API vs Dataset API? - Vodafone

I have one dataframe, which has only one column "Id". Which is the output of following code=> df2 = df1.groupBy().show() - Vodafone

What are the optimisation techniques you have used? - Vodafone

How to access HBASE table from spark? -wipro

If we have 100gb and 10mb file, How do you load and process the 10 MB file in spark? -wipro

You have 100gb in one table and 1gb of small table. How do you join them in spark? -wipro

How to read JDBC data using spark? -Virtusa

What is the difference between coalesce and repartition? -Virtusa

Explain Optimisation techniques used in hive? -Virtusa

Difference between callable and runnable in Java? -Virtusa

How to delete duplicate email entries from the table where id is primary key? -Virtusa

How does Broadcast variable work in Spark? -HCL

How do you set memory persistent level in Spark? -HCL

How to read a cassandra table into spark dataframe? -HCL

How to split a column in a dataframe into multiple columns? -HCL

What is DAG? -wipro

How spark behaves internally when you do spark-submit. -wipro

What is Lazy Evaluation? -wipro

How to split large data frame into halves ? -wipro

How you read data from any external source and save it as a Hive table. -Wipro

How you enable Hive in Spark? -Wipro

How can you read XML file in Spark? -Wipro

Parquet vs avro when to choose which? -Wipro

Optimisation technique in spark (explain, sort merge, shuffle sort, and broadcast join) with scenario? -Wipro

Write any sort algorithm in Scala. -Wipro

Palindrome code in Scala? -Wipro

How will you count the number of records in very large dataframe? -Wipro

Can you return anything in Broadcast Variable? -Wipro

Difference between Broadcast and Persist? -Wipro

How spark decides jobs and stages (explain with transformation) -Abzooba

Difference between Dataset and Dataframe -Abzooba

Write UDF (for a scenario) -Abzooba

Write case when in spark dataframe -Abzooba

Reverse a list -Abzooba

Explain cost based optimizer -Powerupcloud Technologies

one big-query was there in we which we need to find out which one of these are Transformation and actions -Powerupcloud Technologies

Datasets:

Product( product\_ id, product\_ name, product\_ type, product\_ version, product\_price)

Customer (customer\_ id, customer\_ first\_name, customer\_ last\_name, phone\_number)

Sales (transaction\_ id, customer\_ id, product\_id, timestamp, total\_amount, total\_quantity)

a) Display the customer name who made the second most purchases in the month of May 2013.

b) Got 3 valuable customers by total amount for each month. - EPAM

 How do you access data lake in your spark application? -EPAM

 What is DataSet and DF. If you're using how DataSet is useful?

How catalyst optimizer works internally?

How to define partitions and buckets based on cluster size or user choice? -Legato

 How to define windowing functions in spark? -Virtusa

 What is a stage in spark? if there are 10 transformations in a job, how many stages will be created? -Capgemini

 What is accumulator? what is the actual use of it? -Capgemini

 I have a files of 10 GB and 5 gb. Can we use broadcast join in this scenario? - Capgemini

 How do you retrieve 5th row from a file? -Synechron

 How do you remove header from a file? -Synechron

 What are lag, lead and windowing in DataFrames? -Synechron

 There are 100 columns in a dataframe and I need to select 97 columns? How you will do it? - Cognizant

How do you determine driver and executor memory to be specified while running a job? - Cognizant

 What is map side join in spark? -Cognizant

 What are compression techniques in Hadoop? -Reliance Jio

 How to write a spark streaming application where we are taking data from kibana. -Reliance Jio

 How to read picture format in Scala? -Reliance Jio

How to check in sqoop if all the data was loaded properly? -Reliance Jio

 Difference between sparksession and sparkcontext. -Hoonar tekwurks

 Suppose if an RDD is deleted in Spark job, How do we can recover it and what is the backend mechanism of recovering RDD?

How do you achieve broadcast join automatically without doing it manually? And how do you set up your driver program to detect where broadcast join can be good to use.

How do you achieve memory cache in spark? -Accenture

How to find log directory location from ps -ef output?

 How to find allocated heap memory with which namenode is running?

Suppose RAM is 16GB but after 4GB utilization Swap utilization reaches 100%. Explain how it is happening?

How SSL works?

How pagination and memory mapping works in Linux?

What is the difference between rsync and scp?

- Cloudera

What is RDD? What are the multiple ways to create RDD? -NICE SOFTWARE SOLUTIONS

I have multiple files at HDFS location,How do I read them all at once into RDD?  -NICE SOFTWARE SOLUTIONS

I have created a Dataframe. Can I change the data type for a particular column? If yes, please write code for it  -NICE SOFTWARE SOLUTIONS

How to rename the column of already created dataframe?  -NICE SOFTWARE SOLUTIONS

When I am reading data from Dataframe and showing it, I am getting truncated values, how to avoid values getting truncated? -NICE SOFTWARE SOLUTIONS

How spark is better compared to Hadoop mapreduce?  -NICE SOFTWARE SOLUTIONS

What are the disadvantages of spark?  -NICE SOFTWARE SOLUTIONS

Map join vs reducer join differences. -CTS

Write a query to show id and name of latest time records.

How will you eliminate the duplicate records -Paypal

 Static vs dynamic partitioning. -CTS

 Table 1 has 10 records, table 2 has 5 records, result should have 5 records from table 1 that does not exist in table 1 -CTS

 Optimisation techniques - Partitioning, and criteria to choose a column to partition. -CTS

 External vs Internal - at what scenario we choose them, other than table drop reason? -CTS

Table A have 20000 records and Table B have 50 million records how will you handle the joins ?  -Paypal

 ORC vs Parquet -Paypal

Map vs Flatmap -Paypal

 20000 records in a table. copy them to two separate files with records equally splitted into HDFS using Sqoop. How will you handle this, table does not have constraints? -Paypal

Can a partition be dropped from an existing table -CTS

Can a partition be added to an existing table -CTS

ORS vs Parquet, what scenario we choose them -CTS

How will you handle the corrupted data in hive -Altimetric

Write a program to print the possible combinations of any given word

INPUT = “CAT”

Expected Result: “CAT”,“CTA”,“ACT”,“ATC”,“TCA”,“TAC” -Altimetric

What is the difference between Coalesce and Repartition ? -Capgemini

On what principle does Hash Partitioner work in spark ? -Capgemini

How to increase/decrease the number of partitions while submitting spark job? -Capgemini

How to read a csv file into dataframe and write into a Hive table? -Capgemini

How do you convert a dataframe into rdd and then rdd into list? -Altimetric

How to add a new column to a dataframe with a constant value? -Altimetric

What is the best optimization technique in spark and why? -Altimetric

The internal of catalyst optimizer. - Happiest Mind

How to access different versions of s3 bucket in spark? - Happiest Mind

Explain spark architecture or explain what happens if you do spark submit. -  Infosys

Repartition and coalesce (explain with scenario). - Acceldata.io

Will spark be able to read file having more size than allocated Ram? - Acceldata.io

Describe cluster configuration? - Acceldata.io

Serialisation deserialization and GC tuning. - Acceldata.io

Broadcast variable and Accumulator variable with scenario and why used and at which stage of pipeline? - Acceldata.io

Given a time series data which is a clickstream of user activity is stored in hive, ask is to enrich the data with session id.

Session Definition:

- Session expires after inactivity of 30 mins, because of inactivity no clickstream record will be generated.

- Session remains active for a total duration of 2 hours

Steps:

- Local data in Hive table.

- Read the data from Hive using spark batch (scala) to do the computation.

- Save the results in parquet with enriched data.

Note: please do not use direct spark-sql.

timestamp userid

2018-01-01T11:00:00Z u1

2018-01-01T12:00:00Z u1

2018-01-01T11:00:00Z u2

2018-01-02T11:00:00Z u2

2018-01-01T12:15:00Z u1

 - Sapient

In addition to the problem statement given in question 2 assume below scenario as well and design hive table based on it:

- Get Number of sessions generated in a day.

- Total time spent by a user in a day.

- Total time spent by a user over a month.

Here are the guidelines and instructions for the solution of the above queries:

- Design Hive table

- Write script to create the table

- Load data into table

- Write all the queries in Spark SQL

- Sink in the direction of using partitioning, bucketing, etc.

timestamp userid

2018-01-01T11:00:00Z u1

2018-01-01T12:00:00Z u1

2018-01-01T11:00:00Z u2

2018-01-02T11:00:00Z u2

2018-01-01T12:15:00Z u1

- Sapient

Suppose we have following table in hive

Name|age|subject|marks

Ramesh|18|Maths|98

Suresh|18|Physics|95

Amit|17|Maths|87

Ranjan|17|Maths|88

Saikat|18|Physics|84

Avijit|18|Physics|96

Find the student who got the 2nd highest number in a particular subject?

o/p

Ranjan|17|Maths|88

Suresh|18|Physics|95

- TCS

What is the difference between sparksession and sparkcontext? - Iris software

What are the standards which we should keep in mind in order to write optimized code? - Iris software

What is the difference in client and cluster mode? - Iris software

Lets say, I am running a spark job in local more and I have a total of 2GB memory available to my machine and I am trying to read a file of 5GB, will it read the file or throw any exception? - Iris software

In which case scenario we get memory out of exception error while my job is at the verge of completion. - Iris software

What is the difference between reduce and reduceByKey. Which one is action and which is transformation? - Iris software

Why spark is better than Mapreduce? - Capgemini

What is the difference between RDD/Dataframes/Datasets? - Capgemini

What is the use of DAG/RDD Lineage? - Capgemini

Describe Transformation and Actions with example? - Capgemini

Difference between groupByKey & reduceByKey? - Capgemini

Difference between Caching & Persist? - Capgemini

Difference between map vs flatMap? - Capgemini

Difference between sortBy, orderBy, groupBy? - Capgemini

Why overhead memory is used in spark-submit command? -ITC

How to decide for executive memory, cores, driver memory in spark-submit command? -ITC

How to access Hive/beeline from spark.-ITC

Difference between client and cluster mode. -ITC

What is Broadcast variable and what is the maximum size of file which you can broadcast. -ITC

Difference between caching and persist.-ITC

Why spark execution is faster?-ITC

Why we get insufficient memory error in execution?-ITC

Difference between RDD/dataframe/datasets. - CTS

What exactly catalyst optimizer does when triggered a spark submit command?- CTS

How to compare two data frames excluding one columns from both?- CTS

What is lazy collection?- CTS

Explain with example the use of explode()/options()/leteral\_view. -CTC

Why overhead memory is used in spark submit- command? - TCS

How to decide for executor memory, cores, driver memory in spark-submit command? - TCS

Difference between client and cluster mode.- TCS

Can we write an empty file using spark?- TCS

In how many ways we can convert a RDD to dataframe?- TCS

Why compile type safety a beneficial datasets?- TCS

How typecasting security is maintained in datasets?- TCS

Explain the architecture of spark. - Impetus

What is dynamic partition in Spark.- Impetus

What is the benefits MR & spark?- Impetus

What is paired RDD and write a query to get the count of each word in a file?- Impetus

What is transformation & Actions and give few examples.- Impetus

What is RDD lineage?- Impetus

How to recover data in RDD if the flow is broke between?- Impetus

How to run spark over mesos?- Impetus

What is the difference between caching/persist?- Impetus

What is Broadcast variable/ Accumulator?- Impetus

What are partitions in spark? - TCS

What are narrow and wide transformations? - City Bank

What is sparksession?- City Bank

Different formats in hado and ratio between them? - City Bank

What is a w s lambda?- City Bank

Vectorisation and bucketing in hive?- City Bank

Why stages are created in spark?- City Bank

Find the largest consecutive sequence in the list? - Samsung

Write a spark submit skeletal code and explain the terms. - Legato

Which is faster ko coalesce and repartition and why? - IBM

How to write a word count in spark? - TCS

What are the benefits of data sets? - Lowes

How is dataframe  and dataset different from RDD? What optimisation do they have? - Infosys

What are broadcast and accumulators? How do they work? -Infosys

How do you deploy codes in different environment in your project? - Infosys

How to move all distinct records from one table to another? - Infosys

When do we choose mapper = 1 in sqoop? What is its purpose? - Infosys

What is the default number of mappers in sqoop? - Infosys

When and why mapreduce is invoked in hive? - Infosys

Can we use having without groupBy? - Infosys

How do you perform a left outer join on 2 rdd's? - Verizon

How do you perform a right outer join on 2 rdd's? - Verizon

Do you know Spark aggregate functions ? - Verizon

Production functions in spark? - Verizon

Spark accumulators and broadcast variables - Verizon

How can you create data frames? - Verizon

What is the purpose of data frames? - Verizon

What is RDD? How can you improve performance on RDD? - Verizon

Incremental merge in Hive / spark? - Verizon

Spark performance and troubleshooting? - Verizon

what would be the thrift server port on your cluster to connect spark? - Verizon

Difference between groupByKey and reduceByKey? - TCS

What are the actions/ transformations in spark? - TCS

How Spark breaks down tasks given an application submitted?  -TCS

What is the maximum data size you have handled? - TCS

Can Spark to be used to import data from an RDBMS source? - TCS

Replace all null values in a given table with zero. - Legato

Can we execute Hive queries in spark? - Legato

Do you have experience migrating Hive to Spark SQL? What are the issues you faced? - Legato

Explain types of advanced joins in spark. - Legato

In a given table, populate the highest value in "A"column, for each month? - Legato

How do you tune your Spark application for performance? - IBM

Why is Parquet mostly used with Spark rather other formats? Any reason? - IBM

Please write a skeletal spark-submit code and explain how to choose number of executors/cores? - IBM

When to go for broadcast join in Spark? - IBM

What are the file formats Spark supports? - IBM

What is shared variable in spark?  - Lowes

What is the difference between DSM/TDD?  - Lowes

How is fault tolerance achieved in spark?  - Lowes

Reducebykey vs groupbykey difference?  - Lowes

When to use reducebykey or groupbykey?  - Lowes

What are map partitions?  - Lowes

When to use text file() and whole text file () methods?  - Lowes

What parameters you have used in your Spark submit command?  - Alliance Data

How you define executors in Spark submit command?  - Alliance Data

Which is faster dataframe API on Spark SQL? why?  - Alliance Data

I have a dataframe1 where I am doing some groupby and another dataframe2 and I am doing union of both dataframe and then I am writing it some file then how many stages it will create? And while writing the DF in file whether any stage will be created for that or not?  - Alliance Data

Explain The Spark architecture and flow of it? - TCS

 When we should use narrow transformation and when we should prefer to use wide transformation? - TCS

 How you schedule Spark job?-TCS

What is broadcast hash join? - Synechron

 Spark repartition and coalesce.-  Wipro

 Spark jdbc connectivity details. - Wipro

Cache persist.-  Wipro

How data is processed and connected to downstreams? - Wipro

Optimisation techniques in Spark. - Wipro

Difference between DAG and lineage graph. - Wipro

Explain reduceleft and reduce. - Wipro

Explain about reduce fold. - Wipro

Difference between map and flatmap? - Wipro

Broadcast variables and accumulators. -  Wipro

Spark dag flow. - Tetrasoft

Read JSON and XML file in spark. -  Tetrasoft

Slots in spark. -  Tetrasoft

Difference between RDD, dataframe and data set. - Tetrasoft

Incremental loading in spark. -  Tetrasoft

Write a program to find the number of words in a text file. Why should dataframe not rdd? - Knoldus software

Write a program to increase a salary by 10% of employee if he has salary greater than 40k

input:-

Id,name,salary

1,mani,50k

output:-

Id,name,increased\_salary

1,mani,55k

Note: try to do this without using withcolumn option.

 - Knoldus software

In my spark job, there are a lot of joins and hence the job is taking a lot of time to run. What are the optimizations I can use? - Tavant Technologies

How do you decide on the number of executors for a spark? -  job Cap Gemini

What are the optimisations done in your project? - L&T Infotech

Explain the end-to-end pipeline of your project. - Tavant Technologies

Difference between Spark context and Spark session? - Tavant Technologies

What is out of memory error in Spark job? -  Capgemini

Difference between groupByKey and reduceByKey? -  Capgemini

Explain static and dynamic partition pruning in spark? - Capgemini

Explain broadcast variables in Spark?  - Capgemini

Advantage of repartition over coalesce? - Tavant Technologies

Write a spark SQL query to find out the maximum value per group? - Tavant Technologies

What is the size of your cluster? - Tavant Technologies

What is the difference between cache and persist? - Tavant Technologies

Write a spark SQL query to find out the duplicate records in a dataframe? - Tavant Technologies

What are narrow and wide transformations? - Capgemini

What is the difference between transformation and action? - Capgemini

Suppose I am reading from a file, doing some transformation and want to back to the same file in  Spark SQL, I am not able to write it? What could be the reason?  - Capgemini

Suppose I have a could which has an array of integer values. - Capgemini

Not visible

Can we increase partition in coalesce? - Apple

Why does repartition works slow compared with coalesce? - Apple

What is Spark kaise? - Apple

What are transformation and action? - Securonix

What are the spark optimization techniques in spark? - Securonix

What are catalyst and vectorization? - Securonix

Explain how join will be performed between a small table and a large table in spark? - Securonix

What is the difference between temp view and global temp view? - lumiq.ai

How do you create datasets from JSON files? - lumiq.ai

How do you convert existing RDDs to Spark datasets? - lumiq.ai

Is Spark better than mapreduce? why? - lumiq.ai

What exactly is DAG, what will be the vertices and edges? - lumiq.ai

If you do so many replication, what will happen? - lumiq.ai

Memory issue in Spark how you will deal with it? - lumiq.ai

Why Spark creates DAG? - lumiq.ai

Whenever you submit a spark job, how it internally works? - lumiq.ai

Spark architecture explanation? - Lumiq.ai

If you have a spark DataFrame which has a column with NULL values, Which function do you use to replace NULL values with some string? - Alliance Data

 If you have data as below: - Alliance Data

 What happens in cluster when user invokes SPARK SUBMIT? -Jio

 Write code to import data from RDBMS to Hadoop using spark. - Jio

 How would you repartition your data in Spark? - Paypal

 Given the below 2 data sets find out how much time did an employee stayed in office.

Data Set1

empid Name

1 Animesh

2 Dheeraj

3 Priya

4 Nayra

Data Set2

empid ScanType Time

1 login 8/17/2020 9:00

2 login 8/17/2020 9:05

1 logout 8/17/2020 9:30

1 login 8/17/2020 9:35

3 login 8/17/2020 9:55

2 logout 8/17/2020 13:00

3 logout 8/17/2020 17:05

1 logout 8/17/2020 18:00

Result

empid Name TotalTimeInOffice

1 Animesh 9 hrs

2 Dheeraj 3 hrs 55 min

3 Priya 5 hrs

4 Nayra 0 hrs

 - Paypal

If you have data as below: - Alliance Data

If you have a spark DataFrame which has a column with NULL values, Which function do you use to replace NULL values with some string? - Alliance Data

Brief about your project. architecture. -TCS

What is the source of data? -TCS

What is the file size and format you receive on daily basis? -TCS

What is the type of data --history data or ??-TCS

What operations do you perform after you receive the data? -TCS

What is your upstream? -TCS

What is the DAG? -TCS

Difference between driver and executor? -TCS

Given the below file and my Hive table has "," separator. How you will load the data into Hive?

EmpName|Empid

Raj|1133

- TCS

Map in rdd and map in dataframes, map in Scala collection api -TCS

You have an external table in Hive and given a location of HDFS for a file. If we delete the table then if we have its schema and again create a same table then will you be able to see the data? - Accenture

You have 3 partitions having 100 rows each. You do a count(\*). Then you add one more partition with 100 rows. Then again count\*. What will be the value? - Accenture

How to get number of partitions?- Accenture

Explain the architecture of MAP-REDUCE. What are all the phases in Map-reduce program? - Accenture

Given a program in mapreduce where number of mapper is 3 and number of reducer is 4. Then how many combiner and partitioner will be created? - Accenture

What are higher order functions in Scala? Explain with a use-case or example. - Accenture

Which are all the query optimisation techniques in Hive. - Accenture

Difference between Hive managed and external tables. some use-case related questions. - Accenture

Why we use bucketing? How bucketing internally works? - Capgemini

Is it compulsory to use partitioning if we go for bucketing? Can we use bucketing without partitioning? - Capgemini

What is special about a Avro and Parquet file formats? Use case. -Capgemini

What is avro in hive. explain with use-case. -Capgemini

What happens internally if you do bucketing/partitioning of Hive tables. What will be the no of reducers? -Capgemini

Difference between partitioning and bucketing?-Capgemini

Based on what factors you decide the number of buckets.-Capgemini

What day-to-day operations you do in your work? -Capgemini

How you load your Spark dataframe into Hive external tables?-Capgemini

What are optimisation techniques in Spark? - Harman

What is driver memory and executor memory? - Harman

What issues have you troubleshooted in Spark? - Modak Analytics

What is GC error in Spark how to avoid it? - Modak Analytics

What is difference between mapreduce and spark? - Mindtree

Explain DAG cycle in spark? -Altimetric

What error you faced during last Spark job in production environment? - TCS

What is difference between coalesce and repartition? - TCS

What are the parameters we pass during Spark submit? - TCS

What optimisation techniques you have used in your last project? - TCS

What is difference between rdd and dataframe? - Synechron

How can we join two large data set in spark?- Synechron

How dag stage form for a Spark job?- Synechron

How can we change the number of task during processing of Spark program.- Synechron

Real-time question from self join concept using Spark dataframe. - Mindtree

What is difference between groupbykey and reducebykey? - Mindtree

How to write a table in Hive table using Spark dataframe by reading a text file. - L&T Infotech

What is coalesce function?- L&T Infotech

What is lazy evaluation?- L&T Infotech

How can we determine number of partition in spark?- L&T Infotech

What is difference between Spark and mapreduce?- L&T Infotech

What is shared variable in Spark explain each with example?- L&T Infotech

What is the use of case class in Spark?- L&T Infotech

What if we put less number of columns than the number of parameters of case class? - L&T Infotech

Explain Spark architecture? - Walmart

Write Spark program that includes dynamic filter query. -Walmart

How glom difference from map partition in spark? - Walmart

Perform the below operation. - Walmart

Difference Hadoop and spark, how does Spark achieve fast performance than Hadoop? - Optum

Spark cache ( Difference between cache and persistent) - Optum

Spark lazy evaluation. - Optum

 Performance tuning in spark. - Optum

Difference between rdd and data frame? - Optum

Bucketing in Spark dataframe, pushdown feature.- Optum

Spark SQL create temporary tables - advantages.- Optum

Coalesce, Date, Timestamp on DataFrame, Spark SQL Explode function.- Optum

Spark SQL dataframe self join use case, save as Hive table.- Optum

Connect to external sources like Oracle, teradata using Spark.- Optum

Basic Spark Transformations and Actions using pyspark -- create RDD, map(), flatMap(), filter(), Union(), Collect(), take (n), count (), max(), min() etc

groupByKey vs reduceByKey  - Land O Lakes

Difference between dataframe and RDD, dataframe vs datasets. - Land O Lakes

RDD operation, narrow and wide in transformations, checkpointing. - Land O Lakes

Sparksession and Sparkcontext.- Land O Lakes

Deploy modes and cluster manager for spark.- Land O Lakes

Spark optimisation- physical and logical, how to handle data skews.- Land O Lakes

DAG Scheduler.- Land O Lakes

Steps to process JSON and XML files using spark.- Land O Lakes

Write a spark word count program. - TCS

Difference between repartition and coalesce?- TCS

What is dataFrame?- TCS

What is a dense rank?- TCS

What is RDD?- TCS

You have 6 Machines 16 Cores and 64GB Ram… Explain best memory management. - Ness

If I have 100000 records I want to take only 10,000 records how should I do?- Ness

What is catalyst optimizer?- Ness

What is the difference between DAG and Lineage? - Ness

What is RDD? -Wipro

What is DataFrame? -Wipro

What does leftSemi and left anti join work? - Legato

What are the optimization techniques used in Spark? -Virtusa

What is the difference between cache() and persist()? -Virtusa

How Spark execute a job? -Virtusa

If i try to increase the partition size using coalesce, what will happen? - Kreuzwerker

 If one rdd is of string type in another rdd is of int type, what happens if i try union method? - Kreuzwerker

Groupbykey vs reducebykey. - Neustar Inc

Spark context vs Spark session. - Reliance Jio

How you are helping client from 2 years ( 1 put 2 year experience) - Accenture

Difference between coalesce and repartition? - Accenture

How you check piece of code that you developed from your end, before proceeding to cluster? - Accenture

Which API you used for Apache Spark? - Accenture

What are collections in Scala and how can you find the maximum number in maxArray If val maxArray = Array(10,40,60,05,50)? - Accenture

How reduce works with Array in finding max in it? - Accenture

Difference between companion and singleton explain with a programming example. - Accenture

What are traits and explain with a programming example? - Accenture

If a:Int = 10 how can we cast a into float. - Accenture

How to define a method in Scala? - Accenture

How much data flows into your cluster? - Accenture

How do you check performance issues with Spark job? - Accenture

What measures you took for improving performance? - Accenture

What is broadcast variable and why we use it? -  Accenture

Difference between groupBy and reduceBy, map and mapPartition? -  Accenture

Explain UDF in hive write a simple program on UD

F for subtract and sum? -  Accenture

What are major challenges you faced in project? -  Accenture

How The Spark architecture work? -  Bold Technologies

How the transformation that we have applied and in between Spark application, How Spark will rebuild or call those transformation? -  Bold Technologies

Difference between DAG and DAG scheduler? -  Bold Technologies

How partitionby will distribute the rows? -  Bold Technologies

How Spark code can be optimised? -  Bold Technologies

Explain Spark architecture once spark-submit is done to job is submitted? -  Publicis Sapients

What is encoder in spark? -  Publicis Sapients

Suppose I have 100 columns in a dataframe, I want to select columns only which select columns only which specific string in column (achieve without hardcoding) -  Publicis Sapients

Windows function in Spark Scala? -  Publicis Sapients

Difference between map and mapPartition? -  Publicis Sapients

How will you decide the number of executors and memory to be assigned for the  particular Spark Application? -  Fractal Analytics

How to write the data in Spark with one file -  Fractal Analytics

What all the file formats you have worked with it -  Fractal Analytics

How to handle the small file problem -  Fractal Analytics

Difference between AVRO, ORC and parquet file format  -  Fractal Analytics

Difference between coalesce and repartition? -  NIIT Technologies

Difference between dataframe and datasets? -  NIIT Technologies

How rdd is capable of rebuilding in case of failure? -  NIIT Technologies

Spark architecture? -  NIIT Technologies

What is driver program and executor program? -  NIIT Technologies

How to handle the failures of Spark job? -  NIIT Technologies

How you submit the spark job? -  NIIT Technologies

Project architecture? -  NIIT Technologies

Explode function in dataframe? -  NIIT Technologies

Hive optimization ? -  NIIT Technologies

Difference between external and internal table in hive? -  NIIT Technologies

Static and dynamic partitioning? -  NIIT Technologies

For available data in data warehouse with bucket by date, count the no. Of records for everyday? - Wynk

How Distinct and Count works internally? - Wynk

What is broadcast join? - wynk

Using Dataframe, take sample text file and count the occurrence of word and also reduce the no. of shuffle partition from 200 to 10. - Knoldus software

Using Dataframe, take any sample data and create user defined schema and increase the salary of those employee whose salary is 3 lakh, then increase it by 5% - Knoldus software

What are the parameters passes to the spark-submit? - Telstra

How to reduce shuffling ? - Telstra

What is broadcast variable? - Telstra

How to debug Spark application? - Telstra

Which Spark framework and version used in your project? - Telstra