Tell about Yourself

Why Bigdata has been chosen instead of common RDBMS DB?

How did you arrived at the capacity planning and size of the Cluster?

Why Apache NiFi has been chosen for data ingestion instead of other ingestion tools like Flume, Sqoop etc.,

Any other tool used apart from Apache NiFi for data Ingestion irrespective of data type either structure or unstructured?

Advantage of Kafka and why it has been chosen?

How you arrived the sizing of Kafka?

Can you process structured data in Kafka?

What is Hapi framework?

Why HBase chosen?

How HBase is handling the real-time data (or) what part of HBase architecture achieves the real-time data faster.

Have worked in SPARK?

What is RDD in SPARK?

Why RDD is Immutable?

Have you done SPARK related Performance tuning and explain those?

What is Repartitioning in SPARK and how it happens?

How the plan gets created in the SPARK program?

What is stage in SPARK and how the Order of the execution in SPARK happens with Multiple stages?

Have you done group by in a stage?

Can we do Map and group by in a single transformation?

Have you used SPARK SQL?

Are you aware of Datasets & DataFrames in SPARK?

What is the programming language used in SPARK?

Importance of Final in Java? Especially Final in Variable, Class & Method.

Have you used Data structures in Java?

What are the Data structures implemented in your program?

How the program gets invoked once the user interacts in the UI screen?

What is Pattern in Java

What is Singleton Pattern

How will you do a Java program to get the friend of a friend in Facebook?

what’s your experience in Bigdata?

Can you describe the SPARK project which you worked?

What do you understand by RDD?

What is the advantage of creating RDD?

An RDD of Employee object and each employee have addresses. What operation to be done to an RDD with the object Employee id & Address as a string.

What is the purpose of filter operation?

What is Reduce By key & Group by Key Operation and which one is better?

What are the different modes of launching a SPARK job?

Which cluster manager used?

How many executors you launch for streaming job?

How comfortable with Core Java?

How an object immutable in SPARK?

What is the property makes it immutable?

Do you cache anything in SPARK RDD? Why

Bolts & Nuts of Apache framework for high speed event processing?

Having a big RDD with 1Million records. How to find the sum of all Integer values and how it happens internally?

How to copy set of files from one cluster to another cluster?

What is the storage format of your data?

Advantages of using SPARK?

I have 14 executors, each has 10 GB of memory and I should process 1000GB of data? How it will do the in-memory processing.?

What are the different modes of launching a SPARK job?

Having RDD of Integers, how to bring odd numbers in one RDD and even numbers in another RDD?

Describe the FLAT MAP function.

RDD of Employee id and it has list of employee reports to employee. Create an RDD with Employee id and its corresponding subordinate’s id.

What is Reduce By key & Group by Key?

Reduce by key will shuffle the data?

Do you cache your RDD?

What are the different cache strategy?

How much memory you keep it for executor?

If you run an executor with 1GB memory, how much data it can process?

How many executors you run?

Degree of parallelism?

How long you have been working in SPARK?

Which part of SPARK you are working? SPARK streaming

What is the abstraction used in SPARK Streaming?

What’s your experience in Map Reduce?

How to do the map reduce word count program in SPARK?

How do you print the data in SPARK?

Is there any risk in using Collect method in SPARK?

Difference between Reduce by key & Group by key?

Are you aware of Partitions in SPARK?

What are the transformations to tweak the no. of partitions?

There is a file and filtered in an RDD Scenario, file as input and filter on that. Filtered output1. Bis logic, output 1 to multiple times. What is the best approach?

RDD is stored in memory, why need to do cache again?

Which version of Hive you used?

Which distribution of hadoop you used? Cloudera/EMR…

How would you handle updates in Hive?

If I have a scenario to update the data, how do you handle that scenario?

How would you update a record in Partition level?

What kind of join used in Map reduce? Map side join with distributed cache.

Which scenario, you will choose map side join and reduce side join?

Can I do Map side join for a huge dataset in both the side (left & right)?

If one of the Map reduce job fails, what are the steps you follow to debug?

How to find the log of the file if one of the job failed in a distributed environment?

Can we do retry option in Map reduce job?

Experience of Bigdata in short?

Have you worked in SPARK?

How to merge & sort the two sorted integer arrays and how the code should look like, without collections?

Time complexity of the algorithm?

Having a data file with varchar data, Calculate sum of all the numbers in the file?

How will you submit the SPARK job?

Diff between MAP and Flat Map?

Have you worked on SPARK ML?

What version of SPARK you worked?

Which one is faster RDD or Data frame? And what is the major difference?

What is the Size of Cluster

Optimization in SPARK Code?

What kind of serialization used in SPARK?

How will you integrate Hive and HBASE?