**Previous Infosys Interview Questions (separated by associate)**

How to count the number of times a word appears in a text file?(was required to write code to solve this)   
What is an RDD?  
 How to create a dataframe in spark?   
How to create and RRD?   
How to create the schema of a dataframe in spark?   
What the different type of joins in spark?   
How to do a join?

Interview questions asked during interview: Here is a list of the interview questions that I could recall:  
How do you add a column to a DF?  
How do you rename a column in a DF?  
Difference between an RDD/DS/DF?  
How to check if a column already exists in a DF?  
Explain Hive Structure?  
How to submit a job to Spark and what parameters are needed?  
What is lazy evaluation?  
How to read a csv in Scala?  
How Spark SQL works in the background?  
How HQL works in the background?  
Difference between Drop/Truncate?  
How to do a join in spark?  
Difference between Group By/Sort By?  
How to move a table from Hive to Spark?

What do we use spark-submit for  
How to read a CSV file in Spark  
Write a Kafka Topic  
How did we deal with “off setters” in our project I’m not sure if that was the exact word  
How we use Kafka in our projects  
Describe the most recent project

How to use Spark-submit  
What is my tech stack/ what projects have I worked on  
How to create an RDD  
Transformations and Actions in Spark  
Different Hive table types  
How to load a CSV from HDFS to HIVE Table

Explain a project I worked on like P3  
Reversing a string in scala   
Right Join in SQL   
Declare a RDD   
HDFS commands to load into Hive table and copytoHDFS  
Types of Hive tables   
Transformation vs Actions   
What is an RDD   
What is a map

I was asked:   
What is an RDD  
what did i know about spark  
what is a producer  
what are joins in SQL  
type out the command to add file to HDFS from VM  
what is Hadoop  
what is HDFS  
explain MapReduce  
what kind of tables are there in Hive  
type out how I would count the characters in a string in python

Asked me how I would rename a dataframe's column  
asked me how I would read from a text file  
headers in a csv file  
how I would set the names of columns  
how would I drop a specific column if it exists  
How did I use spark as an analytical engine, (I wasn’t exactly sure what they meant by this)  
Asked difference between Hive and SQL (I interpreted this as meaning a RDBMS)  
Asked about P2 and how we did analysis

They asked me to explain Project 1  
General SQL questions and one query which I forgot.   
Python they asked about the project and how to do joins.   
General Python questions and one synthax question in how to create a table.   
Hive questions on partitions and buckets when to use one or the other.   
Spark they asked about what makes it different from Hive and its advantages.   
NO Kafka questions.

Interview questions asked during interview: Here is a list of the interview questions that I could recall:  
How do you add a column to a DF?  
How do you rename a column in a DF?  
Difference between an RDD/DS/DF?  
How to check if a column already exists in a DF?  
Explain Hive Structure?  
How to submit a job to Spark and what parameters are needed?  
What is lazy evaluation?  
How to read a csv in Scala?  
How Spark SQL works in the background?  
How HQL works in the background?  
Difference between Drop/Truncate?  
How to do a join in spark?  
Difference between Group By/Sort By?  
How to move a table from Hive to Spark?

Tell me about how you used Hadoop  
What is the difference between a Map and a Set  
Spark Context  
Kafka  
What is compression  
What different file types are you familiar with  
Whay other Programing languages are you familiar with?

Describe a little bit about yourself, and your most recent project.  
How would you write and read a JSON in Python? - code  
How would you count the number of occurrences in a string in Python? - code  
Describe Project 1.  
What is HDFS?  
What is the difference between SQL and HDFS (HiveQL)? - Asked for clarification because it was asked unusual.  
What is an RDD?  
What is the difference between an RDD, DataSet, and DataFrame?  
Describe partitioning and bucketing.  
Explain pseudo code for Project 3?  
Explain how AWS was incorporated into Project 3?  
What is S3?  
What is Spark, and how would you use spark-submit?  
What is DAG in Spark?  
What are the components of KAFKA?  
How would you start KAFKA, create a topic, consumer, and producer?

Tell me about yourself   
Tell me about your projects and the technologies you used   
What is HDFS and why do we use it   
How do you load a json file and print its contents using python   
What is spark  
What is an RDD  
What transformations did you use in your projects

Pass by value or pass by reference?  
Explain YARN  
Explain HDFS   
How to print a specific character in a string?  
Partitioning and Bucketing

Explain a little bit about yourself and your last project.  
Difference between and RDD/DF/DS.  
Count the number of times a character appears in a string with python.  
Main components of Kafka.  
What is ‘Zookeeper’ in Kafka?  
Explain pass by value and pass by reference.  
Explain UDF: User Defined Functions.  
Difference between Local File System and HDFS.  
How do you define Technical Architecture for a project?  
Can you explain DAG: Directed Acyclic Graph

**Questions/feedback from most recent cohort:**

Most of the questions she asked me were about Databases in general, Big Data, Spark, Hadoop, and Hive.

Most of my questions were mostly conceptual. I was also asked about OOP. I talked a little bit about creating an abstract class in Python.

I could not tell her how to change a column in a Spark Dataframe. I guessed the answer

Some of the many random questions: What is a reducer function?   
What is the group by and reduce by key in hive?   
How do you use Union in SQL? With example   
Most of the concentration seemed to be on giving SQL examples.   
Emphasis was not on the projects or the introduction of yourself; there didn’t seem to be a lot of interest in that area.

Difference between Aggregates and Scalars  
Difference between Having and Where  
Difference between delete and truncate  
DDL v DML  
Bunch of query questions  
Talk about hive and Hadoop  
Difference between MapReduce and Spark

What are transformations?  
Write an SQL function that will remove duplications from a table  
Write a python function that …  
What is checkpoint?  
Difference between Managed and Unmanaged tables  
Difference between Spark and Hadoop

What are the phases in MapReduce  
What transformations can you perform in python  
What are the benefits of python over other programming languages  
What is bucketing   
What program would you write to bucket in Hive  
A lot of questions in HIVE

What was a DataMart  
What was a slowly changing dimension

**OVERALL TIPS FOR INTERVIREWS**

Review your projects, even parts of them you specifically did not complete. About a third of the cumulative interview time was taken up by questions about the projects (especially p2 and p3)

Be very strong in SQL (Syntax, the different query clauses/keywords, functions, JOINs)

Review Hive (creating tables, partitioning/bucketing, managed vs external, where Hive stores data, how data is stored)

Pros and cons of using specific technologies (MapReduce vs Spark, Hive, SQL)

Basics of Spark (RDD, DF)