Spark Heavy Questions:

1. What is the business case of your pipeline (the problem your company is trying to solve)?

2. What is your architecture of your pipeline(s)?

3. What technology stack have you used, and why?

4. What is your specific contribution within the pipeline?

5. Difference between RDD, data frame, and data set?

6. What is the catalyst optimizer?

7. What is the project tungsten and when do you use it?

8. What's the size of the data?

9. How do you calculate the size of your data?

10. How many nodes are in the cluster?

11. How do you calculate the number of executors and memory inside your cluster?

12. What is data skewness?

13. How do you solve the problem of skewness?

14. What is a broadcast variable?

15. What kind of transformations have you been doing in your data?

16. Difference between groupByKey versus reduceByKey?

17. What is shuffling?

18. How do you decrease shuffling?

19. Common issues with memory in Spark?

20. Difference between coalesce and repartition (or partition)?

21. What is the standard garbarge collector for Spark?

22. How to read a file in Spark?

23. Difference between Spark and MapReduce?

24. What is lazy evaluation?

25. What is a trait?

26. What are case classes and what are they used for?

27. What is the DAG?

28. Define the lineage of a Spark job.

29. How do you submit your job to a Spark cluster?

30. How to create RDDs in Spark?

31. How to convert RDD to Dataframes?

32. What are the partitions in Spark

33. What is an action?

34. What are transformations?

35. What is a pair RDD?

36. What is the Driver?

37. What is an executor?

38. Tell me what are the submissions modes for a Spark job?

39. Type of cluster managers in Spark?

40. What is a sparse vector?

41. What is an accumulator?

42. What are sliding windows?

43. Define caching in Spark.

44. Levels of persistence in Spark?

45. Difference between persistence and cache?

46. What are check points?

47. What is fault tolerance in Spark?

48. Different types of compression for Spark?

49. What is a map, versus a flat map, versus a fold?

50. What is the difference between an input block versus a partition in an RDD?

KAFKA Heavy Questions:

1. What is Kafka?

2. What is a topic?

3. What's a producer? a consumer?

4. What are the brokers?

5. Why do we use the offset?

6. What happens if broker A goes down?

7. What happens if consumer goes does?

8. Explain the role of an offset.

9. Explain the role of Zookeeper and how it interacts with Kafka.

10. What are the partitions in Kafka?

11. Describe your Kafka architecture.

12. Explain the concept of leader and follower.

13. What are ISRs?

14. How do you ensure parallel consumption from Kafka?

15. What's the default retention for Kafka?

16. What is multi-tenancy in Kafka?

17. How do you handle change of schema in your Kafka scheme?

18. Differnce between Confluent and Kafka?

19. What is the log anatomy?

20. How do you tune Kafka?

21. How does Kafka log aggregation?

22. Features of Kafka streams?

23. What are the delivery mechanisms for Kafka?

24. Why is better to code in Java for Kafka?

25. Difference betweem a Broker and a Server?

26. What are serdes?

27. Difference between Flume and Kafka?

28. Difference between Kafka, RabbitMQ, JMS?

29. Explain how you can get exactly one message from Kafka.

30. Why is replication required in Kakfa?

31. Where do you save your offsets?