1. What is the difference between Spark and Hadoop?
2. What are the differences between functional and imperative languages, and why is functional programming
important?
3. What is a resilient distributed dataset (RDD), explain showing diagrams?
4. Explain transformations and actions (in the context of RDDs)
5. What are the Spark use cases?
6. Why do we need transformations? What is lazy evaluation and why is it useful?
7. What is ParallelCollectionRDD?
8. Explain how ReduceByKey and GroupByKey work?
9. What is the common workflow of a Spark program?
10. Explain Spark environment for driver. Ref
11. What are the transformations and actions that you have used in Spark?
12. How can you minimize data transfers when working with Spark?
13. What is a lineage graph?
14. Describe the major libraries that constitute the Spark Ecosystem
15. What are the different file formats that can be used in SparkSql?
16. What are Pair RDDs?
17. What is the difference between persist() and cache()
18. What are the various levels of persistence in Apache Spark?
19. Which Storage Level to choose?
20. Explain advantages and drawbacks of RDD
21. Explain why dataset is preferred over RDDs?
22. How to share data from Spark RDD between two applications?
23. Does Apache Spark provide check pointing?
24. Explain the internal working of caching?
25. What is the function of Block manager?
26. Why does Spark SQL consider the support of indexes unimportant?
27. How to convert existing UDTFs in Hive to Scala functions and use them from Spark SQL?
Explain with example
28. Why use dataframes and datasets when we have RDD?
29. What is a Catalyst and how does it work?
30. What are the top challenges developers faces while writing Spark applications?
31. Explain the difference in implementation between DataFrames and DataSet?
32. How is memory handled in Datasets?
33. What are the limitations of dataset?
34. What are the contentions with memory?
35. Show Command to run Spark in YARN client mode?
36. Show Command to run Spark in YARN cluster mode?
37. What is Standalone and YARN mode?
38. Explain client mode and cluster mode in Spark?
39. Which cluster managers are supported by Spark?
40. What is Executor memory?

41. What is DStream and what is the difference between batch and Dstream in Spark streaming?
42. How does Spark Streaming work?
43. Difference between map() and flatMap()?
44. What is reduce() action, Is there any difference between reduce() and reduceByKey()?
45. What is the disadvantage of reduce() action and how can we overcome this limitation?
46. What are Accumulators and when are accumulators truly reliable?
47. What is Broadcast Variables and what advantage do they provide?
48. What is piping? Demonstrate with an example of a data pipeline.
49. What is a driver?
50. What does a Spark Engine do?
51. What are the steps that occur when you run a Spark application on a cluster?
52. What is a schema RDD/DataFrame?
53. What are Row objects?
54. How does Spark achieve fault tolerance?
55. What parameter is set if cores need to be defined across executors?
56. Name few Spark Master system properties?
57. Define Partitions in reference to Spark implementation?
58. Differences between how Spark and MapReduce manage cluster resources under YARN.
59. What is GraphX and what is PageRank?
60. What does MLlib do?
61. What is a Parquet file?
62. Why is Parquet used for Spark SQL?
63. What is schema evolution and what is its disadvantage, explain schema merging in reference to parquet file? Ref
64. Will Spark replace MapReduce?
65. What is Spark Executor?
66. Name the different types of Cluster Managers in Spark
67. How many ways we can create RDDs, show example?
68. How do you flatten rows in Spark? Explain with example.
69. What is Hive on Spark?
70. Explain Spark Streaming Architecture?
71. What are the types of Transformations on DStreams?
72. What is Receiver in Spark Streaming, and can you build custom receivers?
73. Explain the process of Live streaming storing DStream data to database?
74. How is Spark streaming fault tolerant?
75. Explain transform() method used in dSteam?
76. What file systems does Spark support?
77. How is data security achieved in Spark?
78. Explain Kerberos security?
79. Name the various types of distributing that Spark supports?
80. Show some example queries using the Scala DataFrame API.
81. What are the conditions where Spark driver can parallelize dataSets as RDDs?

82. Can repartition() operation decrease the number of partitions?
83. What is the drawback of repartition() and coalesce() operations?
84. In a join operaton for example val joinVal = rddA.join(rddB) will it generate partition?
85. Consider the following code in Spark, what is the final value in fVal variable?
86. Scala pattern matching - Show various ways code can be written?
87. What is the return result when a query is executed using Spark SQL or HIVE? Hint: RDD or dataframe/dataset?
88. If we want to display just the schema of a dataframe/dataset what method is called?
89. Show various implementations for the following query in Spark?
90. What are the most important factors you want to consider when you start machine learning project?
91. As a data scientist, which algorithm would you suggest if legal aspects and ease of explanation to non technical
people are the main criteria?
92. For the supervised learning algorithm, what percentage of data is split between training and test dataset?
93. Compare performance of Avro and parquet file formats and their usage (in the context of Spark)
94. Spark MAster Exposes a set of REST API's to submit and monito applications. Which data format is used for these web services?
95. When you should not use Spark?
96. Can you use Spark to access and analyze data stored in Cassandra databases?
97. With which mathematical properties can you achieve parallelism?
98. What are various types of Partitioning in Apache Spark?
99. How to set partitioning for data in Apache Spark?