Core Hadoop (HDFS & MapReduce)
1. What is Hadoop, and what are its key components?
2. Explain the architecture of HDFS.
3. What is a NameNode and DataNode in HDFS?
4. How does Hadoop handle fault tolerance?
5. What is a block in HDFS? What is the default size?
6. Explain the read and write operations in HDFS.
7. What is the role of the Secondary NameNode?
8. What is a checkpoint in HDFS?
9. How does Hadoop handle large-scale data processing?
10. What is speculative execution in Hadoop?

MapReduce
11. Explain the MapReduce programming model.
12. What are the phases of a MapReduce job?
13. What is a Combiner in MapReduce?
14. What is the difference between `InputFormat` and `OutputFormat`?
15. What is a `Writable` in Hadoop?
16. How does partitioning work in MapReduce?
17. What is the role of the `JobTracker` and `TaskTracker`?
18. How does Hadoop handle data locality?
19. What is a `SequenceFile` in Hadoop?
20. How can you optimize a MapReduce job?

YARN & Hadoop Ecosystem
21. What is YARN, and how does it work?
22. What is the difference between Hadoop 1 and Hadoop 2?
23. Explain the role of `ResourceManager` and `NodeManager` in YARN.
24. What is Apache Hive, and how does it work with Hadoop?
25. What is the difference between Hive and traditional RDBMS?
26. What is HBase, and how does it differ from HDFS?
27. What is Apache Pig? How does Pig Latin differ from SQL?
28. What is Sqoop, and how is it used in Hadoop?
29. What is Flume, and how does it help in data ingestion?
30. What is Zookeeper’s role in Hadoop?

Advanced Hadoop
31. What is a Hadoop DistCp, and when is it used?
32. How does Hadoop handle small files?
33. What is HDFS Federation?
34. What is Hadoop Rack Awareness?
35. How does Hadoop ensure data security?
36. What is Kerberos, and how does it work with Hadoop?

37. What is HDFS Safe Mode?
38. What is a Hadoop Fair Scheduler?
39. How does Hadoop handle compression?
40. What is Avro, and how is it used in Hadoop?

Performance & Troubleshooting
41. How do you debug a failing MapReduce job?
42. What are common bottlenecks in Hadoop?
43. How do you optimize HDFS storage?
44. What happens when a DataNode fails?
45. How do you handle data skew in MapReduce?
46. What is speculative execution, and when is it useful?
47. How do you monitor Hadoop cluster performance?
48. What are the best practices for Hadoop cluster sizing?
49. How do you handle NameNode failure?
50. What are some alternatives to Hadoop MapReduce?

---

2. Apache Spark (60 Questions)
Spark Core & RDDs
51. What is Apache Spark, and how does it differ from Hadoop?
52. Explain Spark's architecture.
53. What is an RDD? How is it created?
54. What are the key features of Spark?
55. What are transformations and actions in Spark?
56. What is lazy evaluation in Spark?
57. How does Spark handle fault tolerance?
58. What is the difference between `cache()` and `persist()`?
59. What are the different storage levels in Spark?
60. What is a DAG in Spark?

Spark SQL & DataFrames
61. What is Spark SQL? How does it differ from Hive?
62. What is a DataFrame in Spark?
63. How do you convert an RDD to a DataFrame?
64. What is a Catalyst Optimizer?
65. What is a Dataset in Spark?
66. How does Spark SQL handle structured data?
67. What are UDFs in Spark SQL?
68. How do you optimize Spark SQL queries?
69. What is Parquet, and why is it used in Spark?
70. How does Spark integrate with Hive?

Spark Streaming & Structured Streaming
71. What is Spark Streaming?
72. What is a DStream?
73. How does Spark Streaming process real-time data?
74. What is the difference between batch and streaming in Spark?
75. What is checkpointing in Spark Streaming?
76. What is Structured Streaming?
77. How does Structured Streaming differ from Spark Streaming?
78. What are watermarks in Structured Streaming?
79. How does Spark handle late data in streaming?
80. What are some use cases of Spark Streaming?

Spark MLlib & GraphX
81. What is Spark MLlib?
82. How does MLlib differ from traditional ML frameworks?
83. What are Pipelines in Spark MLlib?
84. How do you handle feature extraction in MLlib?
85. What is GraphX, and how is it used?
86. How does Spark handle distributed machine learning?
87. What are some limitations of MLlib?
88. How do you evaluate a model in Spark MLlib?
89. What is Cross-Validation in Spark ML?
90. How does Spark integrate with TensorFlow/PyTorch?

Performance Tuning & Optimization
91. How do you optimize a Spark job?
92. What is shuffling in Spark, and how can you minimize it?
93. What is broadcast join in Spark?
94. How does partitioning improve Spark performance?
95. What is the role of `spark.default.parallelism`?
96. How do you handle data skew in Spark?
97. What are accumulator variables in Spark?
98. How do you monitor Spark jobs?
99. What are common causes of Spark job failures?
100. How does Spark handle memory management?

Cluster Management & Deployment
101. What are the different cluster managers in Spark?
102. How does Spark run on YARN?
103. What is dynamic allocation in Spark?
104. How do you submit a Spark job to a cluster?
105. What is the difference between client and cluster mode in Spark?
106. How do you configure Spark for high availability?
107. What are the best practices for Spark logging?

108. How do you secure a Spark cluster?
109. What is the role of `spark-submit`?
110. How does Spark integrate with Kubernetes?

---

3. Deep Learning Platforms (50 Questions)
TensorFlow & Keras
111. What is TensorFlow, and how does it work?
112. Explain TensorFlow's computational graph.
113. What are Tensors in TensorFlow?
114. What is the difference between TensorFlow 1.x and 2.x?
115. What is Keras, and how does it relate to TensorFlow?
116. How do you define a neural network in TensorFlow?
117. What is `tf.data` API used for?
118. How does TensorFlow handle distributed training?
119. What is a TensorFlow Session? (TF1.x)
120. What is `@tf.function` in TensorFlow?

PyTorch
121. What is PyTorch, and how does it differ from TensorFlow?
122. What are PyTorch Tensors?
123. How does PyTorch handle dynamic computation graphs?
124. What is `torch.nn.Module`?
125. How do you train a model in PyTorch?
126. What is Autograd in PyTorch?
127. How does PyTorch support GPU acceleration?
128. What are DataLoaders in PyTorch?
129. How do you save and load a PyTorch model?
130. What is TorchScript?

Distributed Deep Learning
131. What is Horovod, and how does it work?
132. How does distributed training work in TensorFlow?
133. What is `tf.distribute.Strategy`?
134. How does PyTorch support distributed training?
135. What is a parameter server in deep learning?
136. What is Ring-AllReduce?
137. How does NVIDIA's NCCL help in distributed training?
138. What is Federated Learning?
139. How does DeepSpeed improve distributed training?
140. What is mixed-precision training?

Model Optimization & Deployment

141. What is quantization in deep learning?
142. How do you optimize a TensorFlow model for inference?
143. What is TensorRT?
144. How do you deploy a PyTorch model in production?
145. What is ONNX, and how is it used?
146. What is model pruning?
147. How does transfer learning work in deep learning?
148. What is a frozen graph in TensorFlow?
149. How do you convert a PyTorch model to TensorFlow?
150. What is TensorFlow Serving?

Advanced Topics (Transformers, GANs, etc.)
151. What is a Transformer model?
152. How does BERT work?
153. What are GANs, and how are they trained?
154. What is reinforcement learning in deep learning?
155. How does YOLO work for object detection?
156. What is AutoML?
157. What are Diffusion Models?
158. How does LoRA work in fine-tuning LLMs?
159. What is a Vision Transformer (ViT)?
160. How do you fine-tune an LLM using Hugging Face?

---

4. Big Data & Cloud Integration (40 Questions)
Cloud Platforms (AWS, GCP, Azure)
161. What is Amazon EMR, and how does it work with Hadoop/Spark?
162. How does AWS Glue help in Big Data processing?
163. What is Google Dataproc?
164. How does Azure HDInsight work?
165. What is AWS Athena?
166. How does BigQuery differ from traditional Hadoop?
167. What is Snowflake, and how does it fit into Big Data?
168. How does Delta Lake improve data lakes?
169. What is AWS Kinesis, and how does it compare to Spark Streaming?
170. How do you deploy Spark on Kubernetes?

Data Engineering & ETL
171. What is a data pipeline in Big Data?
172. What is Apache Airflow, and how is it used?
173. How does Kafka integrate with Spark?
174. What is Change Data Capture (CDC) in Big Data?
175. What is a data lake vs. a data warehouse?

176. How do you handle schema evolution in Big Data?
177. What is Apache Beam?
178. How does dbt help in data transformation?
179. What is a Slowly Changing Dimension (SCD) in ETL?
180. How do you ensure data quality in Big Data?

Scalability & Performance
181. What is sharding in distributed databases?
182. How does indexing improve query performance?
183. What is columnar storage, and why is it efficient?
184. How does caching help in Big Data systems?
185. What is a Bloom filter, and how is it used in Big Data?
186. How do you handle backpressure in streaming systems?
187. What is a lambda architecture?
188. What is a kappa architecture?
189. How do you design a scalable data warehouse?
190. What are the best practices for Big Data security?

Emerging Trends
191. What is MLOps, and how does it relate to Big Data?
192. How does AI integrate with Big Data platforms?
193. What is edge computing in Big Data?
194. How does blockchain work with Big Data?
195. What is a data mesh architecture?
196. How does real-time analytics work in modern systems?
197. What is the role of Rust in Big Data processing?
198. How does DuckDB fit into the Big Data ecosystem?
199. What is the future of Hadoop in the cloud era?
200. What are the key challenges in modern Big Data systems?

Certainly! Preparing for interviews in Big Data technologies like **Hadoop**, **Spark**, and **recent deep learning platforms** is crucial for roles in data engineering and data science. Below is a comprehensive list of **200 technical questions** to aid your preparation. These questions are categorized by topic to facilitate focused study.

Hadoop

General Hadoop Concepts

1. What is Hadoop, and what are its core components?
2. Explain the differences between Hadoop 1.x and Hadoop 2.x.
3. What are the main features of Hadoop?
4. How does Hadoop ensure data reliability?
5. What is the role of the NameNode in HDFS?
6. What is the role of the DataNode in HDFS?
7. Explain the concept of a Secondary NameNode.
8. What is the purpose of the JobTracker in MapReduce?
9. What is the role of the TaskTracker in MapReduce?
10. How does Hadoop achieve fault tolerance?

HDFS (Hadoop Distributed File System)

11. What is HDFS, and how does it work?
12. Explain the replication mechanism in HDFS.
13. What are HDFS blocks, and why are they important?
14. How does HDFS handle data integrity?
15. What is the default block size in HDFS?
16. Can you change the replication factor in HDFS? How?
17. What happens when a DataNode fails in HDFS?
18. How does HDFS balance the data across DataNodes?
19. What are the benefits of using HDFS over traditional file systems?
20. How can you access HDFS data?

MapReduce

21. What is MapReduce, and how does it work?
22. Explain the phases of a MapReduce job.
23. What is the role of the combiner in MapReduce?
24. How does the partitioner function in MapReduce?
25. What is shuffling and sorting in MapReduce?
26. How can you optimize a MapReduce job?
27. What are the limitations of MapReduce?
28. How does MapReduce handle task failures?
29. What is speculative execution in MapReduce?
30. How do you monitor and debug a MapReduce job?

YARN (Yet Another Resource Negotiator)

31. What is YARN, and why was it introduced?
32. Explain the architecture of YARN.
33. What are the components of YARN?
34. How does YARN manage resources?
35. What is the role of the ResourceManager in YARN?
36. What is the role of the NodeManager in YARN?
37. How does YARN handle job scheduling?
38. What is the ApplicationMaster in YARN?
39. How does YARN support multi-tenancy?
40. What are the benefits of using YARN over the older MapReduce framework?

Hadoop Ecosystem

41. What is Apache Hive, and how does it relate to Hadoop?
42. Explain the use of Apache Pig in the Hadoop ecosystem.
43. What is HBase, and how does it integrate with Hadoop?
44. How does Sqoop facilitate data transfer in Hadoop?
45. What is Flume, and when would you use it?
46. Explain the role of Oozie in Hadoop workflows.
47. What is Zookeeper, and why is it important in Hadoop?
48. How does Mahout relate to Hadoop?
49. What is the purpose of Avro in Hadoop?
50. How does Parquet differ from other storage formats in Hadoop?

Apache Spark

General Spark Concepts

51. What is Apache Spark, and how does it differ from Hadoop MapReduce?
52. Explain the key features of Apache Spark.
53. What are the main components of the Spark ecosystem?
54. How does Spark achieve fault tolerance?
55. What is an RDD (Resilient Distributed Dataset)?
56. How are RDDs created in Spark?
57. What is the difference between transformations and actions in Spark?
58. Explain the concept of lazy evaluation in Spark.
59. What are the benefits of using Spark over Hadoop?

60. How does Spark handle real-time data processing?

Spark Architecture

61. Describe the architecture of Spark.
62. What is the role of the Spark Driver?
63. What are Executors in Spark?
64. How does Spark manage cluster resources?
65. What is the DAG (Directed Acyclic Graph) in Spark?
66. How does Spark execute a job?
67. What is the role of the Cluster Manager in Spark?
68. How does Spark handle task scheduling?
69. What are the different cluster managers available for Spark?
70. How does Spark ensure data locality?

Spark SQL

71. What is Spark SQL?
72. How does Spark SQL integrate with Spark?
73. What are DataFrames in Spark SQL?
74. How do DataFrames differ from RDDs?
75. What is a Dataset in Spark?
76. How does Spark SQL optimize query execution?
77. What is the Catalyst optimizer in Spark SQL?
78. How can you run SQL queries using Spark?

79. What are the benefits of using Spark SQL over traditional SQL engines?
80. How does Spark SQL handle schema inference?

Spark Streaming

81. What is Spark Streaming?
82. How does Spark Streaming process data?
83. What are DStreams in Spark Streaming?
84. How does Spark Streaming achieve fault tolerance?
85. What is the difference between Spark Streaming and Structured Streaming?
86. How can you handle window operations in Spark Streaming?
87. What are the sources from which Spark Streaming can ingest data?
88. How does backpressure work in Spark Streaming?
89. What is checkpointing in Spark Streaming?
90. How can you ensure exactly-once processing in Spark Streaming?

MLlib (Machine Learning Library)

91. What is MLlib in Spark?
92. What are the key features of MLlib?
93. How does MLlib support machine learning algorithms?
94. What is the difference between MLlib and ML packages in Spark?
95. How can you perform feature extraction using MLlib?
96. What are Pipelines in MLlib?
97. How does MLlib handle model evaluation?

98. What is the role of DataFrames in MLlib?

99. How can you save and load models in MLlib?

100. What are the limitations of MLlib?

GraphX

101. What is GraphX in Spark?

102. How does GraphX represent graphs?

103. What are the key features of GraphX?

104. How can you perform graph computations using GraphX?

105. What is the Pregel API in GraphX?

106. How does GraphX optimize graph processing?

107. What are some common use cases for GraphX?

108. How does GraphX integrate with other Spark components?

109. What is the difference between GraphX and other graph processing frameworks?

110. How can you visualize graphs processed with GraphX?

Deep Learning Platforms

General Deep Learning Concepts

111. What is deep learning, and how does it differ from traditional machine learning?

112. Explain the structure of a neural network.

113. What are activation functions, and why are they important?

114. What is backpropagation in neural networks?

115. How do you prevent overfitting in deep learning models?

116. What is the role of dropout in neural networks?

117. Explain the concept of batch normalization.

118.