**HADOOP Multiple Choice Questions and Answers  :-** HADOOP Interview Questions and Answers pdf free download

1. What does commodity Hardware in Hadoop world mean? ( D )  
  
a) Very cheap hardware  
  
b) Industry standard hardware  
  
c) Discarded hardware  
  
d) Low specifications Industry grade hardware  
  
2. Which of the following are NOT big data problem(s)? ( D)  
  
a) Parsing 5 MB XML file every 5 minutes  
  
b) Processing IPL tweet sentiments  
  
c) Processing online bank transactions  
  
d) both (a) and (c)  
  
3. What does “Velocity” in Big Data mean? ( D)  
  
a) Speed of input data generation  
  
b) Speed of individual machine processors  
  
c) Speed of ONLY storing data  
  
d) Speed of storing and processing data  
  
4. The term Big Data first originated from: ( C )  
  
a) Stock Markets Domain  
  
b) Banking and Finance Domain  
  
c) Genomics and Astronomy Domain  
  
d) Social Media Domain  
  
5. Which of the following Batch Processing instance is NOT an example of ( D)  
  
BigData Batch Processing?  
  
a) Processing 10 GB sales data every 6 hours  
  
b) Processing flights sensor data  
  
c) Web crawling app  
  
d) Trending topic analysis of tweets for last 15 minutes  
  
6. Which of the following are example(s) of Real Time Big Data Processing? ( D)  
  
a) Complex Event Processing (CEP) platforms  
  
b) Stock market data analysis  
  
c) Bank fraud transactions detection  
  
d) both (a) and (c)  
  
7. Sliding window operations typically fall in the category (C ) of\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.  
  
a) OLTP Transactions  
  
b) Big Data Batch Processing  
  
c) Big Data Real Time Processing  
  
d) Small Batch Processing  
  
8. What is HBase used as? (A )  
  
a) Tool for Random and Fast Read/Write operations in Hadoop  
  
b) Faster Read only query engine in Hadoop  
  
c) MapReduce alternative in Hadoop  
  
d) Fast MapReduce layer in Hadoop  
  
9. What is Hive used as? (D )  
  
a) Hadoop query engine  
  
b) MapReduce wrapper  
  
c) Hadoop SQL interface  
  
d) All of the above  
  
10. Which of the following are NOT true for Hadoop? (D)  
  
a) It’s a tool for Big Data analysis  
  
b) It supports structured and unstructured data analysis  
  
c) It aims for vertical scaling out/in scenarios  
  
d) Both (a) and (c)  
  
11. Which of the following are the core components of Hadoop? ( D)  
  
a) HDFS  
  
b) Map Reduce  
  
c) HBase  
  
d) Both (a) and (b)  
  
12. Hadoop is open source. ( B)  
  
a) ALWAYS True  
  
b) True only for Apache Hadoop  
  
c) True only for Apache and Cloudera Hadoop  
  
d) ALWAYS False  
  
13. Hive can be used for real time queries. ( B )  
  
a) TRUE  
  
b) FALSE  
  
c) True if data set is small  
  
d) True for some distributions  
  
14. What is the default HDFS block size? ( D )  
  
a) 32 MB  
  
b) 64 KB  
  
c) 128 KB  
  
d) 64 MB  
  
15. What is the default HDFS replication factor? ( C)  
  
a) 4  
  
b) 1  
  
c) 3  
  
d) 2  
  
16. Which of the following is NOT a type of metadata in NameNode? ( C)  
  
a) List of files  
  
b) Block locations of files  
  
c) No. of file records  
  
d) File access control information  
  
17. Which of the following is/are correct? (D )  
  
a) NameNode is the SPOF in Hadoop 1.x  
  
b) NameNode is the SPOF in Hadoop 2.x  
  
c) NameNode keeps the image of the file system also  
  
d) Both (a) and (c)  
  
18. The mechanism used to create replica in HDFS is\_\_\_\_\_\_\_\_\_\_\_\_. ( C)  
  
a) Gossip protocol  
  
b) Replicate protocol  
  
c) HDFS protocol  
  
d) Store and Forward protocol  
  
19. NameNode tries to keep the first copy of data nearest to the client machine. ( C)  
  
a) ALWAYS true  
  
b) ALWAYS False  
  
c) True if the client machine is the part of the cluster  
  
d) True if the client machine is not the part of the cluster  
  
20. HDFS data blocks can be read in parallel. ( A )  
  
a) TRUE  
  
b) FALSE  
  
21. Where is HDFS replication factor controlled? ( D)  
  
a) mapred-site.xml  
  
b) yarn-site.xml  
  
c) core-site.xml  
  
d) hdfs-site.xml  
  
22. Read the statement and select the correct option: ( B)  
  
It is necessary to default all the properties in Hadoop config files.  
  
a) True  
  
b) False  
  
23. Which of the following Hadoop config files is used to define the heap size? (C )  
  
a) hdfs-site.xml  
  
b) core-site.xml  
  
c) hadoop-env.sh  
  
d) Slaves  
  
24. Which of the following is not a valid Hadoop config file? ( B)  
  
a) mapred-site.xml  
  
b) hadoop-site.xml  
  
c) core-site.xml  
  
d) Masters  
  
25. Read the statement:  
  
NameNodes are usually high storage machines in the clusters. ( B)  
  
a) True  
  
b) False  
  
c) Depends on cluster size  
  
d) True if co-located with Job tracker  
  
26. From the options listed below, select the suitable data sources for flume. ( D)  
  
a) Publicly open web sites  
  
b) Local data folders  
  
c) Remote web servers  
  
d) Both (a) and (c)  
  
27. Read the statement and select the correct options: ( A)  
  
distcp command ALWAYS needs fully qualified hdfs paths.  
  
a) True  
  
b) False  
  
c) True, if source and destination are in same cluster  
  
d) False, if source and destination are in same cluster  
  
28. Which of following statement(s) are true about distcp command? (A)  
  
a) It invokes MapReduce in background  
  
b) It invokes MapReduce if source and destination are in same cluster  
  
c) It can’t copy data from local folder to hdfs folder  
  
d) You can’t overwrite the files through distcp command  
  
29. Which of the following is NOT the component of Flume? (B)  
  
a) Sink  
  
b) Database  
  
c) Source  
  
d) Channel  
  
30. Which of the following is the correct sequence of MapReduce flow? ( C )  
  
f) Map ◊ Reduce ◊ Combine  
  
a) Combine ◊ Reduce ◊ Map  
  
b) Map ◊ Combine ◊ Reduce  
  
c) Reduce ◊ Combine ◊ Map  
  
31 .Which of the following can be used to control the number of part files ( B) in a map reduce program output directory?  
  
a) Number of Mappers  
  
b) Number of Reducers  
  
c) Counter  
  
d) Partitioner  
  
32. Which of the following operations can’t use Reducer as combiner also? (D)  
  
a) Group by Minimum  
  
b) Group by Maximum  
  
c) Group by Count  
  
d) Group by Average  
  
33. Which of the following is/are true about combiners? (D)  
  
a) Combiners can be used for mapper only job  
  
b) Combiners can be used for any Map Reduce operation  
  
c) Mappers can be used as a combiner class  
  
d) Combiners are primarily aimed to improve Map Reduce performance  
  
e) Combiners can’t be applied for associative operations  
  
34. Reduce side join is useful for (A)  
  
a) Very large datasets  
  
b) Very small data sets  
  
c) One small and other big data sets  
  
d) One big and other small datasets  
  
35. Distributed Cache can be used in (D)  
  
a) Mapper phase only  
  
b) Reducer phase only  
  
c) In either phase, but not on both sides simultaneously  
  
d) In either phase  
  
36. Counters persist the data on hard disk. (B)  
  
a) True  
  
b) False  
  
37. What is optimal size of a file for distributed cache? (C)  
  
a) <=10 MB  
  
b) >=250 MB  
  
c) <=100 MB  
  
d) <=35 MB  
  
38. Number of mappers is decided by the (D)  
  
a) Mappers specified by the programmer  
  
b) Available Mapper slots  
  
c) Available heap memory  
  
d) Input Splits  
  
e) Input Format  
  
39. Which of the following type of joins can be performed in Reduce side join operation? (E)  
  
a) Equi Join  
  
b) Left Outer Join  
  
c) Right Outer Join  
  
d) Full Outer Join  
  
e) All of the above  
  
40. What should be an upper limit for counters of a Map Reduce job? (D)  
  
a) ~5s  
  
b) ~15  
  
c) ~150  
  
d) ~50  
  
41. Which of the following class is responsible for converting inputs to key-value (c) Pairs of Map Reduce  
  
a) FileInputFormat  
  
b) InputSplit  
  
c) RecordReader  
  
d) Mapper  
  
42. Which of the following writables can be used to know value from a mapper/reducer? (C)  
  
a) Text  
  
b) IntWritable  
  
c) Nullwritable  
  
d) String  
  
43. Distributed cache files can’t be accessed in Reducer. (B)  
  
a) True  
  
b) False  
  
44. Only one distributed cache file can be used in a Map Reduce job. (B)  
  
a) True  
  
b) False  
  
45. A Map reduce job can be written in: (D)  
  
a) Java  
  
b) Ruby  
  
c) Python  
  
d) Any Language which can read from input stream  
  
46. Pig is a: (B)  
  
a) Programming Language  
  
b) Data Flow Language  
  
c) Query Language  
  
d) Database  
  
47. Pig is good for: (E)  
  
a) Data Factory operations  
  
b) Data Warehouse operations  
  
c) Implementing complex SQLs  
  
d) Creating multiple datasets from a single large dataset  
  
e) Both (a) and (d)  
  
48. Pig can be used for real-time data updates. (B)  
  
a) True  
  
b) False  
  
49. Pig jobs have the same run time as the native Map Reduce jobs. (B)  
  
a) True  
  
b) False  
  
50. Which of the following is the correct representation to access ‘’Skill” from the (A)  
  
Bag {‘Skills’,55, (‘Skill’, ‘Speed’), {2, (‘San’, ‘Mateo’)}}  
  
a) $3.$1  
  
b) $3.$0  
  
c) $2.$0  
  
d) $2.$1

***HADOOP Interview Questions and Answers pdf ::***

51. Replicated joins are useful for dealing with data skew. (B)  
  
a) True  
  
b) False  
  
52. Maximum size allowed for small dataset in replicated join is: (C)  
  
a) 10KB  
  
b) 10 MB  
  
c) 100 MB  
  
d) 500 MB  
  
53. Parameters could be passed to Pig scripts from: (E)  
  
a) Parent Pig Scripts  
  
b) Shell Script  
  
c) Command Line  
  
d) Configuration File  
  
e) All the above except (a)  
  
54. The schema of a relation can be examined through: (B)  
  
a) ILLUSTRATE  
  
b) DESCRIBE  
  
c) DUMP  
  
d) EXPLAIN  
  
55. DUMP Statement writes the output in a file. (B)  
  
a) True  
  
b) False  
  
56. Data can be supplied to PigUnit tests from: (C)  
  
a) HDFS Location  
  
b) Within Program  
  
c) Both (a) and (b)  
  
d) None of the above  
  
57. Which of the following constructs are valid Pig Control Structures? (D)  
  
a) If-else  
  
b) For Loop  
  
c) Until Loop  
  
d) None of the above  
  
58. Which of following is the return data type of Filter UDF? (C)  
  
a) String  
  
b) Integer  
  
c) Boolean  
  
d) None of the above  
  
59. UDFs can be applied only in FOREACH statements in Pig. (A)  
  
a) True  
  
b) False  
  
60. Which of the following are not possible in Hive? (E)  
  
a) Creating Tables  
  
b) Creating Indexes  
  
c) Creating Synonym  
  
d) Writing Update Statements  
  
e) Both (c) and (d)  
  
61. Who will initiate the mapper? (A)  
  
a) Task tracker  
  
b) Job tracker  
  
c) Combiner  
  
d) Reducer  
  
62. Categorize the following to the following datatype  
  
a) JSON files – Semi-structured  
  
b) Word Docs , PDF Files , Text files – Unstructured  
  
c) Email body – Unstructured  
  
d) Data from enterprise systems (DB, CRM) – Structured  
  
63. Which of the following are the Big Data Solutions Candidates? (E)  
  
a) Processing 1.5 TB data everyday  
  
b) Processing 30 minutes Flight sensor data  
  
c) Interconnecting 50K data points (approx. 1 MB input file)  
  
d) Processing User clicks on a website  
  
e) All of the above  
  
64. Hadoop is a framework that allows the distributed processing of: (C)  
  
a) Small Data Sets  
  
b) Semi-Large Data Sets  
  
c) Large Data Sets  
  
d) Large and Small Data sets  
  
65. Where does Sqoop ingest data from? (B) & (D)  
  
a) Linux File Directory  
  
b) Oracle  
  
c) HBase  
  
d) MySQL  
  
e) MongoDB  
  
66. Identify the batch processing scenarios from following: (C) & (E)  
  
a) Sliding Window Averages Job  
  
b) Facebook Comments Processing Job  
  
c) Inventory Dynamic Pricing Job  
  
d) Fraudulent Transaction Identification Job  
  
e) Financial Forecasting Job  
  
67. Which of the following is not true about Name Node? (B)& (C) &(D)  
  
a) It is the Master Machine of the Cluster  
  
b) It is Name Node that can store user data  
  
c) Name Node is a storage heavy machine  
  
d) Name Node can be replaced by any Data Node Machine  
  
68. Which of the following are NOT metadata items? (E)  
  
a) List of HDFS files  
  
b) HDFS block locations  
  
c) Replication factor of files  
  
d) Access Rights  
  
e) File Records distribution  
  
69. What decides number of Mappers for a MapReduce job? (C)  
  
a) File Location  
  
b) mapred.map.tasks parameter  
  
c) Input file size  
  
d) Input Splits  
  
70. Name Node monitors block replication process ( B)  
  
a) TRUE  
  
b) FALSE  
  
c) Depends on file type  
  
71. Which of the following are true for Hadoop Pseudo Distributed Mode? (C)  
  
a) It runs on multiple machines  
  
b) Runs on multiple machines without any daemons  
  
c) Runs on Single Machine with all daemons  
  
d) Runs on Single Machine without all daemons  
  
72. Which of following statement(s) are correct? ( C)  
  
a) Master and slaves files are optional in Hadoop 2.x  
  
b) Master file has list of all name nodes  
  
c) Core-site has hdfs and MapReduce related common properties  
  
d) hdfs-site file is now deprecated in Hadoop 2.x  
  
73. Which of the following is true for Hive? ( C)  
  
a) Hive is the database of Hadoop  
  
b) Hive supports schema checking  
  
c) Hive doesn’t allow row level updates  
  
d) Hive can replace an OLTP system  
  
74. Which of the following is the highest level of Data Model in Hive? (c)  
  
a) Table  
  
b) View  
  
c) Database  
  
d) Partitions  
  
75. Hive queries response time is in order of (C)  
  
a) Hours at least  
  
b) Minutes at least  
  
c) Seconds at least  
  
d) Milliseconds at least  
  
76. Managed tables in Hive: (D)  
  
a) Can load the data only from HDFS  
  
b) Can load the data only from local file system  
  
c) Are useful for enterprise wide data  
  
d) Are Managed by Hive for their data and metadata  
  
77. Partitioned tables in Hive: (D)  
  
a) Are aimed to increase the performance of the queries  
  
b) Modify the underlying HDFS structure  
  
c) Are not useful if the filter columns for query are different from the partition columns  
  
d) All of the above  
  
78. Hive UDFs can only be written in Java ( B )  
  
a) True  
  
b) False  
  
79. Hive can load the data from: ( D )  
  
a) Local File system  
  
b) HDFS File system  
  
c) Output of a Pig Job  
  
d) All of the above  
  
80. HBase is a key/value store. Specifically it is: ( E )  
  
a) Sparse  
  
b) Sorted Map  
  
c) Distributed  
  
d) Consistent  
  
e) Multi- dimensional  
  
81. Which of the following is the outer most part of HBase data model ( A )  
  
a) Database  
  
b) Table  
  
c) Row key  
  
d) Column family  
  
82. Which of the following is/are true? (A & D)  
  
a) HBase table has fixed number of Column families  
  
b) HBase table has fixed number of Columns  
  
c) HBase doesn’t allow row level updates  
  
d) HBase access HDFS data  
  
83. Data can be loaded in HBase from Pig using ( D )  
  
a) PigStorage  
  
b) SqoopStorage  
  
c) BinStorage  
  
d) HbaseStorage  
  
84. Sqoop can load the data in HBase (A)  
  
a) True  
  
b) False  
  
85. Which of the following APIs can be used for exploring HBase tables? (D)  
  
a) HBaseDescriptor  
  
b) HBaseAdmin  
  
c) Configuration  
  
d) HTable  
  
86. Which of the following tables in HBase holds the region to key mapping? (B)  
  
a) ROOT  
  
b) .META.  
  
c) MAP  
  
d) REGIONS  
  
87. What is the data type of version in HBase? (B)  
  
a) INT  
  
b) LONG  
  
c) STRING  
  
d) DATE  
  
88. What is the data type of row key in HBase? (D)  
  
a) INT  
  
b) STRING  
  
c) BYTE  
  
d) BYTE[]  
  
89. HBase first reads the data from (B)  
  
a) Block Cache  
  
b) Memstore  
  
c) HFile  
  
d) WAL  
  
90. The High availability of Namenode is achieved in HDFS2.x using (C)  
  
a) Polled Edit Logs  
  
b) Synchronized Edit Logs  
  
c) Shared Edit Logs  
  
d) Edit Logs Replacement  
  
91. The application master monitors all Map Reduce applications in the cluster (B)  
  
a) True  
  
b) False  
  
92. HDFS Federation is useful for the cluster size of: (C)  
  
a) >500 nodes  
  
b) >900 nodes  
  
c) > 5000 nodes  
  
d) > 3500 nodes  
  
93. Hive managed tables stores the data in (C)  
  
a) Local Linux path  
  
b) Any HDFS path  
  
c) HDFS warehouse path  
  
d) None of the above  
  
94. On dropping managed tables, Hive: (C)  
  
a) Retains data, but deletes metadata  
  
b) Retains metadata, but deletes data  
  
c) Drops both, data and metadata  
  
d) Retains both, data and metadata  
  
95. Managed tables don’t allow loading data from other tables. (B)  
  
a) True  
  
b) False  
  
96. External tables can load the data from warehouse Hive directory. (A)  
  
a) True  
  
b) False  
  
97. On dropping external tables, Hive: (A)  
  
a) Retains data, but deletes metadata  
  
b) Retains metadata, but deletes data  
  
c) Drops both, data and metadata  
  
d) Retains both, data and metadata  
  
98. Partitioned tables can’t load the data from normal (partitioned) tables (B)  
  
a) True  
  
b) False  
  
99. The partitioned columns in Hive tables are (B)  
  
a) Physically present and can be accessed  
  
b) Physically absent but can be accessed  
  
c) Physically present but can’t be accessed  
  
d) Physically absent and can’t be accessed  
  
100. Hive data models represent (C)  
  
a) Table in Metastore DB  
  
b) Table in HDFS  
  
c) Directories in HDFS  
  
d) None of the above  
  
101. When is the earliest point at which the reduce method of a given Reducer can be called?  
  
A. As soon as at least one mapper has finished processing its input split.  
  
B. As soon as a mapper has emitted at least one record.  
  
C. Not until all mappers have finished processing all records.  
  
D. It depends on the InputFormat used for the job.  
  
Answer: C  
  
**Explanation:**  
  
In a MapReduce job reducers do not start executing the reduce method until the all Map jobs have completed. Reducers start copying intermediate key-value pairs from the mappers as soon as they are available. The programmer defined reduce method is called only after all the mappers have finished.  
  
Note: The reduce phase has 3 steps: shuffle, sort, and reduce. Shuffle is where the data is collected by the reducer from each mapper. This can happen while mappers are generating data since it is only a data transfer. On the other hand, sort and reduce can only start once all the mappers are done.  
  
Why is starting the reducers early a good thing? Because it spreads out the data transfer from the mappers to the reducers over time, which is a good thing if your network is the bottleneck.  
  
Why is starting the reducers early a bad thing? Because they “hog up” reduce slots while only copying data. Another job that starts later that will actually use the reduce slots now can’t use them.  
  
We can customize when the reducers startup by changing the default value of mapred.reduce.slowstart.completed.maps in mapred-site.xml. A value of 1.00 will wait for all the mappers to finish before starting the reducers. A value of 0.0 will start the reducers right away. A value of 0.5 will start the reducers when half of the mappers are complete. You can also change mapred.reduce.slowstart.completed.maps on a job-by-job basis.  
  
Typically, keep mapred.reduce.slowstart.completed.maps above 0.9 if the system ever has multiple jobs running at once. This way the job doesn’t hog up reducers when they aren’t doing anything but copying data. If we have only one job running at a time, doing 0.1 would probably be appropriate.  
  
102. Which describes how a client reads a file from HDFS?  
  
A. The client queries the NameNode for the block location(s). The NameNode returns the block location(s) to the client. The client reads the data directory off the DataNode(s).  
  
B. The client queries all DataNodes in parallel. The DataNode that contains the requested data responds directly to the client. The client reads the data directly off the DataNode.  
  
C. The client contacts the NameNode for the block location(s). The NameNode then queries the DataNodes for block locations. The DataNodes respond to the NameNode, and the NameNode redirects the client to the DataNode that holds the requested data block(s). The client then reads the data directly off the DataNode.  
  
D. The client contacts the NameNode for the block location(s). The NameNode contacts the DataNode that holds the requested data block. Data is transferred from the DataNode to the NameNode, and then from the NameNode to the client.  
  
Answer: C  
  
 103. When You are developing a combiner that takes as input Text keys, IntWritable values, and emits Text keys, IntWritable values. Which interface should your class implement?  
  
A. Combiner <Text, IntWritable,Text, IntWritable>  
  
A. Reducer <Text, IntWritable,Text, IntWritable>  
  
A. Combiner <Text,Text, IntWritable, IntWritable>  
  
A. Combiner <Text, Text, IntWritable, IntWritable>  
  
Answer: B  
  
104. Indentify the utility that allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer?  
  
A. Oozie  
  
B. Sqoop  
  
C. Flume  
  
D. Hadoop Streaming  
  
E. mapred  
  
Answer: D  
  
105. How are keys and values presented and passed to the reducers during a standard sort and shuffle phase of MapReduce?  
  
A. Keys are presented to reducer in sorted order; values for a given key are not sorted. B. Keys are presented to reducer in sorted order; values for a given key are sorted in ascending order.  
  
C. Keys are presented to a reducer in random order; values for a given key are not sorted.  
  
D. Keys are presented to a reducer in random order; values for a given key are sorted in ascending order.  
  
Answer: A  
  
  
106. Assuming default settings, which best describes the order of data provided to a reducer’s reduce method  
  
A. The keys given to a reducer aren’t in a predictable order, but the values associated with those keys always are.  
  
B. Both the keys and values passed to a reducer always appear in sorted order.  
  
C. Neither keys nor values are in any predictable order.  
  
D. The keys given to a reducer are in sorted order but the values associated with each key are in no predictable order  
  
Answer: D