(1)What is Difference between Secondary namenode, Checkpoint namenode & backupnod Secondary Namenode, a poorly named component of hadoop.  
  
(2)What are the Side Data Distribution Techniques.  
  
(3)What is shuffleing in mapreduce?  
  
(4)What is partitioning?  
  
(5)Can we change the file cached by Distributed Cache  
  
(6)What if job tracker machine is down?  
  
(7)Can we deploy job tracker other than name node?  
  
(8)What are the four modules that make up the Apache Hadoop framework?  
  
(9)Which modes can Hadoop be run in? List a few features for each mode.  
  
(10)Where are Hadoop’s configuration files located?  
  
(11)List Hadoop’s three configuration files.  
  
  
(12)What are “slaves” and “masters” in Hadoop?  
  
  
(13)How many datanodes can run on a single Hadoop cluster?  
  
  
(14)What is job tracker in Hadoop?  
  
  
(15)How many job tracker processes can run on a single Hadoop cluster?  
  
  
(16)What sorts of actions does the job tracker process perform?  
  
  
(17)How does job tracker schedule a job for the task tracker?  
  
  
(18)What does the mapred.job.tracker command do?  
  
  
(19)What is “PID”?  
  
  
(20)What is “jps”?  
  
  
(21)Is there another way to check whether Namenode is working?  
  
  
(22)How would you restart Namenode?  
  
  
(23)What is “fsck”?  
  
  
(24)What is a “map” in Hadoop?  
  
  
(25)What is a “reducer” in Hadoop?  
  
  
(26)What are the parameters of mappers and reducers?  
  
  
(27)Is it possible to rename the output file, and if so, how?  
  
  
(28)List the network requirements for using Hadoop.  
  
  
(29)Which port does SSH work on?  
  
  
(30)What is streaming in Hadoop?  
  
  
(31)What is the difference between Input Split and an HDFS Block?  
  
  
(32)What does the file hadoop-metrics.properties do?  
  
  
(33)Name the most common Input Formats defined in Hadoop? Which one  is default?  
  
  
(34)What is the difference between TextInputFormat and KeyValueInputFormat class?  
  
(35)What is InputSplit in Hadoop?  
  
(36)How is the splitting of file invoked in Hadoop framework  
  
(37)Consider case scenario: In M/R system,  
  
  
- HDFS block size is 64 MB  
  
  
- Input format is FileInputFormat  
  
  
 - We have 3 files of size 64K, 65Mb and 127Mb  
  
  
(38)How many input splits will be made by Hadoop framework?  
  
  
(39)What is the purpose of RecordReader in Hadoop?  
  
  
(39)After the Map phase finishes, the Hadoop framework does “Partitioning, Shuffle and sort”. Explain what happens in this phase?  
  
  
(40)If no custom partitioner is defined in Hadoop then how is data partitioned before it is sent to the reducer?  
  
  
(41)What is JobTracker?  
  
  
(42)What are some typical functions of Job Tracker?  
  
  
(43)What is TaskTracker?  
  
  
(44)What is the relationship between Jobs and Tasks in Hadoop?  
  
  
(46)Suppose Hadoop spawned 100 tasks for a job and one of the task failed. What will Hadoop do?  
  
  
(47)Hadoop achieves parallelism by dividing the tasks across many nodes, it is possible for a few slow nodes to rate-limit the rest of the program and slow down the program. What mechanism Hadoop provides to combat this?  
  
  
(48)How does speculative execution work in Hadoop?    
  
  
(49)Using command line in Linux, how will you  
  
  
- See all jobs running in the Hadoop cluster  
  
  
- Kill a job?  
  
  
(50)What is Hadoop Streaming?    
  
  
(51)What is the characteristic of streaming API that makes it flexible run MapReduce jobs in languages like Perl, Ruby, Awk etc.?  
  
  
(52)What is Distributed Cache in Hadoop?  
  
  
(53)Is it possible to provide multiple input to Hadoop? If yes then how can you give multiple directories as input to the Hadoop job?  
  
  
(54)Is it possible to have Hadoop job output in multiple directories? If yes, how?  
  
  
(55)What will a Hadoop job do if you try to run it with an output directory that is already present? Will it  
  
  
- Overwrite it  
  
  
- Warn you and continue  
  
  
- Throw an exception and exit  
  
  
(56)How can you set an arbitrary number of mappers to be created for a job in Hadoop?  
  
  
(57)How can you set an arbitrary number of Reducers to be created for a job in Hadoop?  
  
  
(58)How will you write a custom partitioner for a Hadoop job?  
  
  
(59)How did you debug your Hadoop code?  
  
  
(60)What is BIG DATA?  
  
  
(61)Can you give some examples of Big Data?  
  
  
(62)Can you give a detailed overview about the Big Data being generated by Facebook?  
  
  
(63)According to IBM, what are the three characteristics of Big Data?  
  
  
(64)How Big is ‘Big Data’?  
  
  
(65)How analysis of Big Data is useful for organizations?  
  
  
(66)Who are ‘Data Scientists’?  
  
  
(67)What are some of the characteristics of Hadoop framework?  
  
  
(68)Give a brief overview of Hadoop history.  
  
  
(69)Give examples of some companies that are using Hadoop structure?  
  
  
(70)What is the basic difference between traditional RDBMS and Hadoop?  
  
  
(71)What is structured and unstructured data?  
  
  
(72)What are the core components of Hadoop?  
  
  
(73)What is HDFS?  
  
  
(74)What are the key features of HDFS?  
  
  
(75)What is Fault Tolerance?  
  
  
(76)Replication causes data redundancy then why is is pursued in HDFS?  
  
  
(77)Since the data is replicated thrice in HDFS, does it mean that any calculation done on one node will also be replicated on the other two?  
  
  
(78)What is throughput? How does HDFS get a good throughput?  
  
  
(79)What is streaming access?  
  
  
(80)What is a commodity hardware? Does commodity hardware include RAM?  
  
  
(81)What is a metadata?  
  
  
(82)Why do we use HDFS for applications having large data sets and not when there are lot of small files?  
  
  
(83)What is a daemon?  
  
  
(84)Is Namenode machine same as datanode machine as in terms of hardware?  
  
  
(85)What is a heartbeat in HDFS?  
  
  
(86)Are Namenode and job tracker on the same host?  
  
  
(87)What is a ‘block’ in HDFS?  
  
  
(88)What are the benefits of block transfer?  
  
  
(89)If we want to copy 10 blocks from one machine to another, but another machine can copy only 8.5 blocks, can the blocks be broken at the time of replication?  
  
  
(90)How indexing is done in HDFS?  
  
  
(91)If a data Node is full how it’s identified?  
  
  
(92)If datanodes increase, then do we need to upgrade Namenode?  
  
  
(93)Are job tracker and task trackers present in separate machines?  
  
  
(94)When we send a data to a node, do we allow settling in time, before sending another data to that node?  
  
  
(95)Does hadoop always require digital data to process?  
  
  
(96)On what basis Namenode will decide which datanode to write on?  
  
  
(97)Doesn’t Google have its very own version of DFS?  
  
  
(98)Who is a ‘user’ in HDFS?  
  
  
(99)Is client the end user in HDFS?  
  
  
(100)What is the communication channel between client and namenode/datanode?  
  
  
(101)What is a rack?  
  
  
(102)On what basis data will be stored on a rack?  
  
  
(103)Do we need to place 2nd and 3rd data in rack 2 only?  
  
  
(104)What if rack 2 and datanode fails?  
  
  
(105)What is a Secondary Namenode? Is it a substitute to the Namenode?  
  
  
(106)What is the difference between Gen1 and Gen2 Hadoop with regards to the Namenode?  
  
  
(107)What is ‘Key value pair’ in HDFS?  
  
  
(108)What is the difference between MapReduce engine and HDFS cluster?  
  
  
(109)Is map like a pointer?  
  
  
(110)Do we require two servers for the Namenode and the datanodes?  
  
  
(111)Why are the number of splits equal to the number of maps?  
  
  
(112)Is a job split into maps?  
  
  
(113)Which are the two types of ‘writes’ in HDFS?  
  
  
(114)Why ‘Reading‘ is done in parallel and ‘Writing‘ is not in HDFS?  
  
  
(115)Can Hadoop be compared to NOSQL database like Cassandra?  
  
  
(116)How can I install Cloudera VM in my system?  
  
  
(117)What is a Task Tracker in Hadoop? How many instances of Task Tracker run on a hadoop cluster  
  
  
(118)What are the four basic parameters of a mapper?   
  
  
(119)What is the input type/format in MapReduce by default?   
  
  
(120)Can we do online transactions(OLTP) using Hadoop? SRVMTrainings   
  
  
(121)Explain how HDFS communicates with Linux native file system   
  
  
(122)What is a JobTracker in Hadoop? How many instances of JobTracker run on a Hadoop Cluster?  
  
  
(123)What is the InputFormat ?   
  
  
(124)What is the InputSplit in map reduce ?   
  
  
(125)What is a IdentityMapper and IdentityReducer in MapReduce ?  
  
  
(126)How JobTracker schedules a task?  
  
  
(127)When is the reducers are started in a MapReduce job?  
  
  
(128)On What concept the Hadoop framework works?   
  
  
(129)What is a DataNode? How many instances of DataNode run on a Hadoop Cluster?   
  
  
(130)What other technologies have you used in hadoop sta ck?   
  
  
(131)How NameNode Handles data node failures?   
  
  
(132)How many Daemon processes run on a Hadoop system?   
  
  
(133)What is configuration of a typical slave node on Hadoop cluster?  
(134) How many JVMs run on a slave node?   
  
  
(135)How will you make changes to the default configuration files?   
  
  
(136)Can I set the number of reducers to zero?   
  
  
(137)Whats the default port that jobtrackers listens ?   
  
  
(138)unable to read options file while i tried to import data from mysql to hdfs. Narendra   
  
  
(139)What problems have you faced when you are working on Hadoop code?   
  
  
(140)how would you modify that solution to only count the number of unique words in all the documents?   
  
  
(141)What is the difference between a Hadoop  and Relational Database and Nosql?   
  
  
(142)How the HDFS Blocks are replicated?   
  
  
(143)What is a Task instance in Hadoop? Where does it run?  
  
  
(144)what is meaning Replication factor?   
  
  
(145)If reducers do not start before all mappers finish then why does the progress on MapReduce job shows something like Map(50%) Reduce(10%)? Why reducers progress percentage is displayed when mapper is not finished yet?  
  
  
(146)How the Client communicates with HDFS?   
  
(147)Which object can be used to get the progress of a particular job   
  
(148)What is next step after Mapper or MapTask?   
  
(149)What are the default configuration files that are used in Hadoop?  
  
(150)Does MapReduce programming model provide a way for reducers to communicate with each other? In a MapReduce job can a reducer communicate with another reducer?  
  
(151)What is HDFS Block size? How is it different from traditional file system block size?  
  
(152)what is SPF?   
  
(153)Where do you specify the Mapper Implementation?   
  
(154)What is a NameNode? How many instances of NameNode run on a Hadoop Cluster?  
  
(155)Explain the core methods of the Reducer?  
  
(156)What is Hadoop framework?  
  
(157)Is it possible to provide multiple input to Hadoop? If yes then how can you give multiple directories as input to the Hadoop job  
  
  
(158)How would you tackle counting words in several text documents?   
  
  
(159)How does master slave architecture in the Hadoop?   
  
  
(160)How would you tackle calculating the number of unique visitors for each hour by mining a huge Apache log? You can use post processing on the output of the MapReduce job.   
  
  
(161)How did you debug your Hadoop code ?   
  
  
(162)How will you write a custom partitioner for a Hadoop job?   
  
  
(163)How can you add the arbitrary key-value pairs in your mapper?  
  
  
(164)what is a datanode?   
  
  
(165)What are combiners? When should I use a combiner in my MapReduce Job?   
  
  
(166)How Mapper is instantiated in a running job?  
  
  
(167)Which interface needs to be implemented to create Mapper and Reducer for the Hadoop?   
  
  
(168)What happens if you don?t override the Mapper methods and keep them as it is?   
  
  
(169)How does an Hadoop application look like or their basic components?  
  
(170)What is the meaning of speculative execution in Hadoop? Why is it important?   
  
  
(170)What are the restriction to the key and value class ?  
  
  
  
  
(171)Explain the WordCount implementation via Hadoop framework ?  
  
  
(172)What Mapper does?   
  
  
(173)what is MAP REDUCE?   
  
  
  
  
(174)Explain the Reducer?s Sort phase?   
  
  
(175)What are the primary phases of the Reducer?   
  
  
(176)Explain the Reducer's reduce phase?   
  
  
(177)Explain the shuffle?   
  
  
(178)What happens if number of reducers are 0?   
  
  
(179)How many Reducers should be configured?   
  
  
(180)What is Writable & WritableComparable interface?   
  
  
(181)What is the Hadoop MapReduce API contract for a key and value Class?   
  
  
(182)Where is the Mapper Output (intermediate kay-value data) stored ?   
  
  
(183)What is the difference between HDFS and NAS ?   
  
  
(184)Whats is Distributed Cache in Hadoop   
  
  
(185)Have you ever used Counters in Hadoop. Give us an example scenario?   
  
  
(186)can we write map reduce program in other than java programming language. how.  
  
  
(187)What alternate way does HDFS provides to recover data in case a Namenode, without backup, fails and cannot be recovered?   
  
  
(188)What is the use of Context object?   
  
  
(189)What is the Reducer used for?   
  
  
(190)What is the use of Combiner?   
  
(191)Explain how input and output data format of the Hadoop framework?  
  
  
(192)What is compute and Storage nodes?   
  
  
(193)what is namenode?   
  
  
(194)How does Mappers run() method works?  
  
  
(195)what is the default replication factor in HDFS?   
  
  
(196)It can be possible that a Job has 0 reducers?   
  
  
(197)How many maps are there in a particular Job?   
  
  
(198)How many instances of JobTracker can run on a Hadoop Cluser?   
  
  
(199)How can we control particular key should go in a specific reducer?   
  
  
(200)what is the typical block size of an HDFS block?   
  
  
(201)What do you understand about Object Oriented Programming (OOP)? Use Java examples.   
  
  
(202)What are the main differences between versions 1.5 and version 1.6 of Java?   
  
  
(203)Describe what happens to a MapReduce job from submission to output?   
  
  
(204)What mechanism does Hadoop framework provides to synchronize changes made in Distribution Cache during runtime of the application   
  
  
(205)Did you ever built a production process in Hadoop ? If yes then what was the process when your hadoop job fails due to any reason   
  
  
(206)Did you ever ran into a lop sided job that resulted in out of memory error, if yes then how did you handled it   
  
  
(207)What is HDFS ? How it is different from traditional file systems?   
  
  
(208)What is the benifit of Distributed cache, why can we just have the file in HDFS and have the application read it   
  
  
(209)How JobTracker schedules a task?  
  
  
(210)How many Daemon processes run on a Hadoop system?  
  
  
(211)What is configuration of a typical slave node on Hadoop cluster? How many JVMs run on a slave node?  
  
  
(212)What is configuration of a typical slave node on Hadoop cluster? How many JVMs run on a slave node?  
  
  
(213)What is the difference between HDFS and NAS ?  
  
  
(214)How NameNode Handles data node failures?  
  
  
(215)Does MapReduce programming model provide a way for reducers to communicate with each other? In a MapReduce job can a reducer communicate with another reducer?  
  
  
(216)Where is the Mapper Output (intermediate kay-value data) stored ?  
  
  
(217)What are combiners? When should I use a combiner in my MapReduce Job?  
  
  
(218)What is a IdentityMapper and IdentityReducer in MapReduce ?  
  
  
(219)When is the reducers are started in a MapReduce job?  
  
  
(220)If reducers do not start before all mappers finish then why does the progress on MapReduce job shows something like Map(50%) Reduce(10%)? Why reducers progress percentage is displayed when mapper is not finished yet?  
  
  
(221)What is HDFS Block size? How is it different from traditional file system block size?  
  
  
(222)How the Client communicates with HDFS?  
  
  
(223)What is NoSQL?  
  
  
(224)We have already SQL then Why NoSQL?   
  
  
(225)What is the difference between SQL and NoSQL?  
  
  
(226)Is NoSQL follow relational DB model?  
  
  
(227)Why would NoSQL be better than using a SQL Database? And how much better is it?  
  
  
(228)What do you understand by Standalone (or local) mode?  
  
  
(229)What is Pseudo-distributed mode?  
  
  
(230)What does /var/hadoop/pids do?  
  
  
(231)Pig for Hadoop - Give some points?  
  
  
  
(232)Hive for Hadoop - Give some points?  
  
  
(233)File permissions in HDFS?  
  
  
(234)what is ODBC and JDBC connectivity in Hive?  
  
  
(235)What is Derby database?  
  
  
(236)What is Schema on Read and Schema on Write?  
  
  
(237)What infrastructure do we need to process 100 TB data using Hadoop?  
  
  
(238)What is Internal and External table in Hive?  
  
  
(239)what is Small File Problem in Hadoop  
  
  
(240)How does a client read/write data in HDFS?  
  
  
(241)What should be the ideal replication factor in Hadoop?  
  
  
(242)What is the optimal block size in HDFS?  
  
  
(243)explain Metadata in Namenode  
  
  
(244)how to enable recycle bin or trash in Hadoop  
  
  
(245)what is difference between int and intwritable  
  
  
(246)How to change Replication Factor (For below cases):  
  
  
(247)In Map Reduce why map write output to Local Disk instead of HDFS?  
  
  
(248)Rack awareness of Namenode  
  
  
(249)Hadoop the definitive guide (2nd edition) pdf  
  
  
(250)What is bucketing in Hive?  
  
  
(251)What is Clustring in Hive?  
  
  
(252)What type of data we should put in Distributed Cache? When to put the data in DC? How much volume we should put in?  
  
  
(253)What is Distributed Cache?  
  
  
(254)What is Partioner in hadoop? Where does it run,mapper or reducer?  
  
  
(255) what are mapreduce new and old apis while writing map reduce program . explain how it works  
  
  
(256)How to write a Custom Key Class?  
  
  
(257)What is the utility of using Writable Comparable (Custom Class) in Map Reduce code?  
  
  
(258)What are Input Format, Input Split & Record Reader and what they do?  
  
  
(259)Why we use IntWritable instead of Int? Why we use LongWritable instead of Long?  
  
  
(260)How to enable Recycle bin in Hadoop?  
  
  
(261)If data is present in HDFS and RF is defined, then how can we change Replication Factor?  
  
  
(262)How we can change Replication factor when Data is on the fly?  
  
  
(262)mkdir: org.apache.hadoop.hdfs.server.namenode.SafeModeException: Cannot create directory /user/hadoop/inpdata. Name node is in safemode.  
  
  
(263)What Hadoop Does in Safe Mode  
  
  
(264)What should be the ideal replication factor in Hadoop Cluster?  
  
  
(265)Heartbeat for Hadoop  
  
  
(266)What will be the consideration while we do Hardware Planning for Master in Hadoop architecture?  
  
  
(267)When should be hadoop archive create  
  
  
(268)what factors the block size takes before creation?  
  
  
(269)In which location Name Node sores its Metadata and why?  
  
  
(270)Should we use RAID in Hadoop or not?  
  
  
(271)How blocks are distributed among all data nodes for a particular chunk of data?  
  
  
(272)How to enable Trash/Recycle Bin in Hadoop?  
  
  
(273)what is hadoop archive  
  
  
(274)How to create hadoop archive  
  
  
(275)How we can take Hadoop out of Safe Mode  
  
  
(276)What is safe mode in Hadoop?  
  
  
(277)Why Mapreduce output written in local disk  
  
  
(278)When Hadoop Enter in Safe Mode  
  
  
(279)Data node block size in HDFS, why 64MB?  
  
  
(280)What is the Non DFS Used  
  
  
(281)Virtual Box & Ubuntu Installation  
  
  
(282)What is Rack awareness?  
  
  
(283)On what basis name node distribute blocks across the data nodes?  
  
  
(284)What is Output Format in hadoop?  
  
  
(285)How to write data in Hbase using flume?  
  
  
(286)What is difference between memory channel and file channel in flume?  
  
  
(287)How to create table in hive for a json input file.  
  
  
(288)What is speculative execution in Hadoop?  
  
  
(289)What is a Record Reader in hadoop?  
  
  
(290)How to resolve the following error while running a query in hive: Error in metadata: Cannot validate serde  
  
  
(291)What is difference between internal and external tables in hive?  
  
  
(292)What is Bucketing and Clustering in Hive?  
  
  
(293)How to enable/configure the compression of map output data in hadoop?  
  
  
(294)What is InputFormat in hadoop?  
  
  
(295)How to configure hadoop to reuse JVM for mappers?  
  
  
(296)What is difference between split and block in hadoop?  
  
  
(297)What is Input Split in hadoop?  
  
  
(298)How can one write custom record reader?  
  
  
(299)What is balancer? How to run a cluster balancing utility?  
  
  
(300)What is version-id mismatch error in hadoop?  
  
  
(301)How to handle bad records during parsing?  
  
  
(302)What is identity mapper and reducer? In which cases can we use them?  
  
  
(303)What is Reduce only jobs?  
  
  
(304)What is crontab? Explain with suitable example.  
  
  
(305)Safe-mode execeptions  
  
  
(306)What is the meaning of the term "non-DFS used" in Hadoop web-console?  
  
  
(307)What is AMI  
  
  
(308)Can we submit the mapreduce job from slave node?  
  
  
(309)How to resolve small file problem in hdfs?  
  
  
(310)How to overwrite an existing output file during execution of mapreduce jobs?  
  
  
(311)What is difference between reducer and combiner?  
  
  
(311)What do you understand from Node redundancy and is it exist in hadoop cluster   
  
  
(312)how to proceed to write your first mapreducer program.  
  
  
(313)How to change replication factor of files already stored in HDFS  
  
  
(314)  How to resolve IOException: Cannot create directory, while formatting namenode in hadoop.  
  
  
(315)How can one set space quota in Hadoop (HDFS) directory  
  
  
(316)How can one increase replication factor to a desired value in Hadoop?