**1.You suspect that your NameNode is incorrectly configured, and is swapping memory to disk. Which Linux commands help you to identify whether swapping is occurring?**

**2.Assuming a cluster running HDFS, MapReduce version 2 (MRv2) on YARN with all settings at their default, what do you need to do when adding a new slave node to cluster?**

**3.You are running a Hadoop cluster with a NameNode on host mynamenode. What are two ways to determine available HDFS space in your cluster?**

**4.You are running a Hadoop cluster with MapReduce version 2 (MRv2) on YARN. You consistently see that MapReduce map tasks on your cluster are running slowly because of excessive garbage collection of JVM, how do you increase JVM heap size property to 3GB to optimize performance?**

**5.You want to understand more about how users browse your public website. For example, you want to know which pages they visit prior to placing an order. You have a server farm of 200 web servers hosting your website. Which is the most efficient process to gather these web server across logs into your Hadoop cluster analysis?**

**6.You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?**

**7. You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?**

**8.You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?**

**9.You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?**

**10. You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?**

**You have A 20 node Hadoop cluster, with 18 slave nodes and 2 master nodes running HDFS High Availability (HA). You want to minimize the chance of data loss in your cluster. What should you do?**

**11 You have installed a cluster HDFS and MapReduce version 2 (MRv2) on YARN. You have no dfs.hosts entry(ies) in your hdfs-site.xml configuration file. You configure a new worker node by setting fs.default.name in its configuration files to point to the NameNode on your cluster, and you start the DataNode daemon on that worker node. What do you have to do on the cluster to allow the worker node to join, and start sorting HDFS blocks?**

**12 Your cluster is configured with HDFS and MapReduce version 2 (MRv2) on YARN. What is the result when you execute: hadoop jar SampleJar MyClass on a client machine?**

**13. Your cluster is configured with HDFS and MapReduce version 2 (MRv2) on YARN. What is the result when you execute: hadoop jar SampleJar MyClass on a client machine?**

**14. Your cluster is configured with HDFS and MapReduce version 2 (MRv2) on YARN. What is the result when you execute: hadoop jar SampleJar MyClass on a client machine?**

**15.Your Hadoop cluster contains nodes in three racks. You have not configured the dfs.hosts property in the NameNode’s configuration file. What results?**

**16. You are migrating a cluster from MApReduce version 1 (MRv1) to MapReduce version 2 (MRv2) on YARN. You want to maintain your MRv1 TaskTracker slot capacities when you migrate. What should you do/**

**17.Choose three reasons why should you run the HDFS balancer periodically?**

**18.You want to node to only swap Hadoop daemon data from RAM to disk when absolutely necessary. What should you do?**

**19.You have recently converted your Hadoop cluster from a MapReduce 1 (MRv1) architecture to MapReduce 2 (MRv2) on YARN architecture. Your developers are accustomed to specifying map and reduce tasks (resource allocation) tasks when they run jobs: A developer wants to know how specify to reduce tasks when a specific job runs. Which method should you tell that developers to implement?**

**20.Assume you have a file named foo.txt in your local directory. You issue the following three commands:  
Hadoop fs –mkdir input  
Hadoop fs –put foo.txt input/foo.txt  
Hadoop fs –put foo.txt input  
What happens when you issue the third command?**

**21.Which YARN process run as “container 0” of a submitted job and is responsible for resource qrequests?**

**22.Which two are features of Hadoop’s rack topology?**

**23.You’re upgrading a Hadoop cluster from HDFS and MapReduce version 1 (MRv1) to one running HDFS and MapReduce version 2 (MRv2) on YARN. You want to set and enforce version 1 (MRv1) to one running HDFS and MapReduce version 2 (MRv2) on YARN. You want to set and enforce a block size of 128MB for all new files written to the cluster after upgrade. What should you do?**

**24.Table schemas in Hive are:**

**25.Which YARN daemon or service monitors a Controller’s per-application resource using (e.g., memory CPU)?**

**26. You have a cluster running with the fair Scheduler enabled. There are currently no jobs running on the cluster, and you submit a job A, so that only job A is running on the cluster. A while later, you submit Job B. now Job A and Job B are running on the cluster at the same time. How will the Fair Scheduler handle these two jobs**

**27.You are configuring your cluster to run HDFS and MapReducer v2 (MRv2) on YARN. Which two daemons needs to be installed on your cluster’s master nodes?**

**28.You are running a Hadoop cluster with a NameNode on host mynamenode, a secondary NameNode on host mysecondarynamenode and several DataNodes.  
Which best describes how you determine when the last checkpoint happened?.**

**29.Which two features does Kerberos security add to a Hadoop cluster?**

**30.Which command does Hadoop offer to discover missing or corrupt HDFS data?**

**31.On a cluster running MapReduce v2 (MRv2) on YARN, a MapReduce job is given a directory of 10 plain text files as its input directory. Each file is made up of 3 HDFS blocks. How many Mappers will run?**

**32.Your cluster has the following characteristics:  
A rack aware topology is configured and on Replication is set to 3 Cluster block size is set to 64MB  
Which describes the file read process when a client application connects into the cluster and requests a 50MB file?**

**33.What does CDH packaging do on install to facilitate Kerberos security setup?**

**34.On a cluster running CDH 5.0 or above, you use the hadoop fs –put command to write a 300MB file into a previously empty directory using an HDFS block size of 64 MB. Just after this command has finished writing 200 MB of this file, what would another use see when they look in directory?**

**35.Your cluster’s mapred-start.xml includes the following parameters  
< name >mapreduce.map.memory.mb< /name >  
< value >4096< /value >  
< name >mapreduce.reduce.memory.mb< /name >  
< value>8192  
And any cluster’s yarn-site.xml includes the following parameters  
yarn.nodemanager.vmen-pmen-ration< /name >  
< value >2.1< /value>  
What is the maximum amount of virtual memory allocated for each map task before YARN will kill its Container?**

**36.You have a cluster running with a FIFO scheduler enabled. You submit a large job A to the cluster, which you expect to run for one hour. Then, you submit job B to the cluster, which you expect to run a couple of minutes only.  
You submit both jobs with the same priority.  
Which two best describes how FIFO Scheduler arbitrates the cluster resources for job and its tasks?**

**37.You are planning a Hadoop cluster and considering implementing 10 Gigabit Ethernet as the network fabric. Which workloads benefit the most from faster network fabric?**

**38.You are configuring a server running HDFS, MapReduce version 2 (MRv2) on YARN running Linux. How must you format underlying file system of each DataNode?**

**39.Which scheduler would you deploy to ensure that your cluster allows short jobs to finish within a reasonable time without starting long-running jobs?**

**40.A slave node in your cluster has 4 TB hard drives installed (4 x 2TB). The DataNode is configured to store HDFS blocks on all disks. You set the value of the dfs.datanode.du.reserved parameter to 100 GB. How does this alter HDFS block storage?**

**41.In CDH4 and later, which file contains a serialized form of all the directory and files inodes in the filesystem, giving the NameNode a persistent checkpoint of the filesystem metadata?.**

**42.For each YARN job, the Hadoop framework generates task log file. Where are Hadoop task log files stored?**

**.**

**43.What two processes must you do if you are running a Hadoop cluster with a single NameNode and six DataNodes, and you want to change a configuration parameter so that it affects all six DataNodes.**

**44.Which process instantiates user code, and executes map and reduce tasks on a cluster running MapReduce v2 (MRv2) on YARN?**

**45.Assuming you’re not running HDFS Federation, what is the maximum number of NameNode daemons you should run on your cluster in order to avoid a “split-brain” scenario with your NameNode when running HDFS High Availability (HA) using Quorum-based storage?**

**46.Your cluster implements HDFS High Availability (HA). Your two NameNodes are named nn01 and nn02. What occurs when you execute the command: hdfs haadmin –failover nn01 nn02?**

**47.Your Hadoop cluster is configuring with HDFS and MapReduce version 2 (MRv2) on YARN. Can you configure a worker node to run a NodeManager daemon but not a DataNode daemon and still have a functional cluster?**

**48.You have just run a MapReduce job to filter user messages to only those of a selected geographical region. The output for this job is in a directory named westUsers, located just below your home directory in HDFS. Which command gathers these into a single file on your local file system?**

**49.You need to analyze 60,000,000 images stored in JPEG format, each of which is approximately 25 KB. Because you Hadoop cluster isn’t optimized for storing and processing many small files, you decide to do the following actions:  
1. Group the individual images into a set of larger files  
2. Use the set of larger files as input for a MapReduce job that processes them directly with python using Hadoop streaming.  
Which data serialization system gives the flexibility to do this?**

**Answer:**

|  |  |
| --- | --- |
|  | **HTML** |
|  | **Avro** |
|  | **SequenceFiles** |
|  | **XML** |

**You decide to create a cluster which runs HDFS in High Availability mode with automatic failover, using Quorum Storage. What is the purpose of ZooKeeper in such a configuration?**

**Answer:**

|  |  |
| --- | --- |
|  | **It both keeps track of which NameNode is Active at any given time, and manages the Edits file. Which is a log of changes to the HDFS filesystem** |
|  | **It only keeps track of which NameNode is Active at any given time** |
|  | **It monitors an NFS mount point and reports if the mount point disappears** |
|  | **If only manages the Edits file, which is log of changes to the HDFS filesystem** |

**You observed that the number of spilled records from Map tasks far exceeds the number of map output records. Your child heap size is 1GB and your io.sort.mb value is set to 1000MB. How would you tune your io.sort.mb value to achieve maximum memory to disk I/O ratio?**

**Answer:**

|  |  |
| --- | --- |
|  | **For a 1GB child heap size an io.sort.mb of 128 MB will always maximize memory to disk I/O** |
|  | **Decrease the io.sort.mb value to 0** |
|  | **Increase the io.sort.mb to 1GB** |
|  | **Tune the io.sort.mb value until you observe that the number of spilled records equals (or is as close to equals) the number of map output records** |

**Which three basic configuration parameters must you set to migrate your cluster from MapReduce 1 (MRv1) to MapReduce V2 (MRv2)?**

**Answer:**

|  |  |
| --- | --- |
|  | **Configure the NodeManager to enable MapReduce services on YARN by setting the following property in yarn-site.xml: < name >yarn.nodemanager.hostname< /name > < value >your\_nodeManager\_shuffle< /value >** |
|  | **Configure the NodeManager hostname and enable node services on YARN by setting the following property in yarn-site.xml: < name >yarn.nodemanager.hostname< /name > < value >your\_nodeManager\_hostname< /value >** |
|  | **Configure a default scheduler to run on YARN by setting the following property in mapredsite.xml: < name >mapreduce.jobtracker.taskScheduler< /name > < Value >org.apache.hadoop.mapred.JobQueueTaskScheduler< /value >** |
|  | **Configure the number of map tasks per jon YARN by setting the following property in mapred: < name >mapreduce.job.maps< /name > < value >2< /value >** |

**A user comes to you, complaining that when she attempts to submit a Hadoop job, it fails. There is a Directory in HDFS named /data/input. The Jar is named j.jar, and the driver class is named DriverClass.  
She runs the command:  
Hadoop jar j.jar DriverClass /data/input/data/output  
The error message returned includes the line:  
PriviligedActionException as:training (auth:SIMPLE) cause:org.apache.hadoop.mapreduce.lib.input.invalidInputException:  
Input path does not exist: file:/data/input  
What is the cause of the error?**

**Answer:**

|  |  |
| --- | --- |
|  | **The output directory already exists** |
|  | **The user is not authorized to run the job on the cluster** |
|  | **The name of the driver has been spelled incorrectly on the command line** |
|  | **The directory name is misspelled in HDFS** |

**ou have a Hadoop cluster HDFS, and a gateway machine external to the cluster from which clients submit jobs. What do you need to do in order to run Impala on the cluster and submit jobs from the command line of the gateway machine?**

**Answer:**

|  |  |
| --- | --- |
|  | **Install the impalad daemon statestored daemon, and daemon on each machine in the cluster, and the impala shell on your gateway machine** |
|  | **Install the impalad daemon, the statestored daemon, the catalogd daemon, and the impala shell on your gateway machine** |
|  | **Install the impalad daemon and the impala shell on your gateway machine, and the statestored daemon and catalogd daemon on one of the nodes in the cluster** |
|  | **Install the impalad daemon on each machine in the cluster, the statestored daemon and catalogd daemon on one machine in the cluster, and the impala shell on your gateway machine** |

**You have a Hadoop cluster HDFS, and a gateway machine external to the cluster from which clients submit jobs. What do you need to do in order to run Impala on the cluster and submit jobs from the command line of the gateway machine?**

**Answer:**

|  |  |
| --- | --- |
|  | **Install the impalad daemon statestored daemon, and daemon on each machine in the cluster, and the impala shell on your gateway machine** |
|  | **Install the impalad daemon, the statestored daemon, the catalogd daemon, and the impala shell on your gateway machine** |
|  | **Install the impalad daemon and the impala shell on your gateway machine, and the statestored daemon and catalogd daemon on one of the nodes in the cluster** |
|  | **Install the impalad daemon on each machine in the cluster, the statestored daemon and catalogd daemon on one machine in the cluster, and the impala shell on your gateway machine** |

**Each node in your Hadoop cluster, running YARN, has 64GB memory and 24 cores. Your yarn.site.xml has the following configuration:  
< property >  
< name >yarn.nodemanager.resource.memory-mb  
32768**

**yarn.nodemanager.resource.cpu-vcores< /name >  
< value >12< /value >  
< /property >  
You want YARN to launch no more than 16 containers per node. What should you do?**

**Answer:**

|  |  |
| --- | --- |
|  | **Modify yarn-sites.xml with the following property: < name >yarn.scheduler.minimum-allocation-mb< /name > < value >4096< /value >** |
|  | **Modify yarn-site.xml with the following property: yarn.scheduler.minimum-allocation-mb < value >2048** |
|  | **No action is needed: YARN’s dynamic resource allocation automatically optimizes the node memory and cores** |
|  | **Modify yarn-site.xml with the following property: < name >yarn.nodemanager.resource.cpu-vccores< /name >** |

**Identify two features/issues that YARN is designated to address:**

**Answer:**

|  |  |
| --- | --- |
|  | **Standardize on a single MapReduce API** |
|  | **Reduce complexity of the MapReduce APIs** |
|  | **Single point of failure in the NameNode** |
|  | **Resource pressure on the JobTracker** |