Academy Training

Hadoop Fundamentals Exam

Exam

The test will be comprised of multiple-choice questions.

1. Someone in your data center unplugs a slave node by accident. Users of the cluster notice via the Hadoop Web UI that the cluster size has shrunken and express concerns about data loss and HDFS performance. The replication factor of all the files in the cluster is unchanged from the default of 3. What can you tell the users?

    A. The HDFS filesystem is corrupt until the administrator re-adds the DataNode to the cluster. The warnings associated with the event should be reported.
    B. After identifying the outage, the NameNode will naturally re-replicate the data and there will be no data loss. The administrator can re-add the DataNode at any time. The client can disregard warnings concerned with this event. Data will be under-replicated but will become properly replicated over time.
    C. The NameNode will identify the outage and re-replicate the data when the clients receive connection failures to the DataNode, so the end users can disregard such warnings.
    D. The NameNode will re-replicate the data after the administrator issues a special command. The data is not lost but is under-replicated until the administrator issues this command.

2. You set the value of dfs.block.size to 64MB in hdfs-site.xml on a client machine, but you set the same property to 128MB on your cluster's NameNode. What happens when the client writes a file to HDFS?

    A. A block size of 128MB will be used
    B. A block size of 64MB will be used
    C. The file will be written successfully with a block size of 64MB, but clients attempting to read the file will fail because the NameNode believes the blocks to be 128MB in size
    D. An exception will be thrown when the client attempts to write the file, because the values are different

3. When is the earliest that the reduce() method of any reduce task in a given job called?

    A. As soon as at least one map task has finished processing its complete input split
    B. Not until all map tasks have completed
    C. It depends on the InputFormat used for the job
    D. As soon as a map task emits at least one record

4. Which of the following are considered to be the 'three Vs' of Big Data?

    A. Veracity, volume, viscosity
    B. Volume, velocity, variety
    C. Velocity, variance, volume
    D. Volume, value, variety

5. You have configured your cluster's dfs.hosts property to point to a file on your NameNode listing all the DataNode hosts allowed to join your cluster. You add a new node to the cluster and update dfs.hosts to include the new host. What do you need to do next to ensure the NameNode reads the change?

    A. You should issue the command hadoop dfsadmin -refreshNodes
    B. You should do nothing as the NameNode re-reads the file every three seconds
    C. You must restart all DataNodes, then restart the NameNode
    D. You should issue the command hadoop dfsadmin -nodeUpdate

6. You are working on a project where you need to chain together Mapreduce and Pig jobs. You also need the ability to use forks, decision points, and path joins. Which of the following ecosystem projects allows you to accomplish this?

    A. Oozie
    B. Sqoop
    C. HBase
    D. HUE
    E. ZooKeeper

7. After a file has been written to HDFS, which of the following operations can you perform? (Choose all that are applicable)

    A. You can delete the file
    B. You can overwrite the file by creating a new file with the same name
    C. You can move the file
    D. You can rename the file
    E. You can update the file's contents

8. Which two daemons typically run on each slave node in a Hadoop cluster running MapReduce v2 (MRv2) on YARN? (Choose two)

    A. ZooKeeper
    B. NodeManager
    C. TaskTracker
    D. Secondary NameNode
    E. JobTracker
    F. JournalNode
    G. NameNode
    H. DataNode

9. What happens under YARN if a Mapper on one node hangs while running a MapReduce job?

    A. After a period of time, the NodeManager will mark the Map task attempt as failed and ask the ApplicationMaster to terminate the container for the Map task.
    B. After a period of time, the ResourceManager will mark the Map task attempt as failed and ask the NodeManager to terminate the container for the Map task
    C. The job will immediately fail
    D. After a period of time, the ApplicationMaster will mark the Map task attempt as failed and ask the NodeManager to terminate the container for the Map task

10. You submit a job to a cluster running MapReduce version 1. You have 10 slave nodes in a single rack, each running a TaskTracker and a DataNode, named node1, node2 … node10. You have NOT specified a rack topology script. Your job has a single Reducer which runs on node4. The output file it writes is small enough to fit in a single HDFS block. How does Hadoop handle writing the output file?

    A. Because no rack topology script has been specified, only one replica of the block will be stored. It will be stored on node4.
    B. Each of the replicas of the block will be stored on a different node, but you do not know which nodes will be used.
    C. The three replicas of the block will be stored on node1, node2 and node3.
    D. The job will fail because you have not specified a rack topology script.
    E. The first replica of the block will be stored on node4. The other two replicas will be stored on other nodes.
    F. Because no rack topology script has been specified, only one replica of the block will be stored on any node.

11. Identify the function performed by a Secondary NameNode daemon configured to run with a single NameNode?

    A. It provides an alternative HDFS endpoint when the NameNode is too busy
    B. It performs real-time backups of the NameNode
    C. It acts as a standby NameNode, providing a high availability profile for clients
    D. It combines the fsimage and edits files produced by the NameNode

12. In HDFS, a file is stored with the permission rw-r--r-- within a directory with the permissions rwxr-xr-x. What does this tell you about the file?

    A. The file's existing contents can be modified by the owner, but no-one else
    B. The file cannot be deleted by anyone but the owner
    C. The file cannot be used as input to a MapReduce job
    D. The file cannot be deleted by anyone

13. Your cluster is running a single NameNode. What happens if the NameNode crashes?

    A. HDFS becomes temporarily unavailable until an administrator starts redirecting client requests to the Secondary NameNode
    B. HDFS becomes unavailable until the NameNode is restored
    C. The Secondary NameNode seamlessly takes over and there is no service interruption
    D. HDFS becomes unavailable to new MapReduce jobs, but running jobs will continue until completion

14. Which of the following does the NameNode store in RAM?

    A. The edits log
    B. Filenames and permissions of data blocks
    C. Contents of files in HDFS
    D. Client information

15. A client application creates an HDFS file named foo.txt with a replication factor of 3. Identify which best describes the file access rules in HDFS if the file has a single block that is stored on data nodes A, B and C?

    A. The file can be accessed if at least one of the DataNodes storing the block is available
    B. Each DataNode locks the local file to prohibit concurrent readers and writes of the file
    C. The file will be marked as corrupt if DataNode B fails during the creation of the file
    D. Each DataNode stores a copy of the file in the local file system with the same name as the HDFS file

16. The Hadoop framework provides a mechanism for coping with machine issues such as fault configuration or impending hardware failure. The JobTracker detects that one or a number of machines are performing poorly and starts more copies of a map or reduce task. What is the feature called where all the tasks run simultaneously and the task that finish first are used?

    A. Combiner
    B. Identity Mapper
    C. Identity Reducer
    D. Speculative Execution

17. Which tool is best suited to import a portion of a relational database every day as files into HDFS, and generate Java classes to interact with that imported data?

    A. Hue
    B. Pig
    C. Oozie
    D. Hive
    E. fuse-dfs
    F. Flume
    G. Sqoop

18. Which statement most accurately describes the relationship between MapReduce and Pig?

A. Pig programs rely on MapReduce but are extensible, allowing developers to do special-purpose processing not provided by MapReduce.
B. Pig provides no additional capabilities to MapReduce. Pig programs are executed as MapReduce jobs via the Pig interpreter.
C. Pig provides the additional capability of allowing you to control the flow of multiple MapReduce jobs.
D. Pig provides additional capabilities that allow certain types of data manipulation not possible with MapReduce.

19. A developer has submitted a long-running MapReduce job with wrong data sets. You want to kill the running MapReduce job so that a new job with the correct data sets can be started. What method can be used to terminate the submitted MapReduce job?

A. hadoop datanode –rollback
B. rmadmin –refreshQueues
C. Open a remote terminal to the node running the ApplicationMaster and kill the JVM
D. yarn application –kill <application_id>
E. Use CTRL-C from the terminal where the MapReduce job was started

20. What must you do if you are running a Hadoop cluster with a single NameNode and six DataNodes, and you wish to change the configuration of all DataNodes?

A. You must restart the NameNode daemon to apply the changes to the cluster
B. You must restart all six DataNode daemons to apply the changes
C. You don't need to restart any daemon, as they will pick up changes automatically
D. You must modify the configuration files on your NameNode where the master configuration files reside for all DataNodes

21. You have configured the Fair Scheduler on your Hadoop cluster. You submit job A, so that ONLY job A is running on the cluster. Job A required more task resources than are available simultaneously on the cluster. Later, you submit job B. Now job A and job B are running on the cluster at the same time. Identify two aspects of how the Fair Scheduler will arbitrate cluster resources for these two jobs. (Choose two)

    A. When job A gets submitted, it consumes all the task resources available on the cluster
    B. When job B gets submitted, job A has to finish first, before job B can be scheduled.
    C. When job A gets submitted, it is not allowed to consume all the task resources on the cluster in case another job is submitted later
    D. When job B gets submitted, it will be allocated task resources while job A continues to run with fewer task resources available to it

22. Which of the following is responsible for running a scheduler to determine how resources are allocated?

    A. ApplicationMaster
    B. NodeManager
    C. ResourceManager
    D. NameNode

23. Using Hadoop's default settings, how much data will you be able to store on your Hadoop cluster if it has 12 nodes with 4TB of raw disk space per node allocated to HDFS storage?

    A. Approximately 3TB
    B. Approximately 12TB
    C. Approximately 16TB
    D. Approximately 48TB

24. What occurs when you run a Hadoop job, specifying an output directory joboutput which already exists in HDFS?

    A. An error will occur immediately, because the output directory must not already exist when a MapReduce job commences
    B. An error will occur after the Mappers have completed but before any Reducers begin to run, because the output path must not exist when the Reducers commence
    C. The job will run successfully. Output from the Reducers will be placed in a directory called joboutput-1
    D. The job will run successfully. Output from the Reducers will overwrite the contents of the existing directory

25. How does the NameNode know which DataNodes are currently available on a cluster?

    A. The NameNode sends a broadcast across the network when it first starts, and DataNodes respond
    B. The NameNode broadcasts a heartbeat on the network on a regular basis, and DataNodes respond
    C. DataNodes are listed in the dfs.hosts file. The NameNode uses that as the definitive list of available DataNodes
    D. DataNodes heartbeat in to the master on a regular basis

26. The following steps are taken to determine how resources are allocated when running a job on a Hadoop cluster.

1 - The ApplicationMaster asks the ResourceManager to assign containers for each Map task
2 - The ResourceManager Scheduler allocates a container for the ApplicationMaster
3 - The ResourceManager asks a NodeManager to launch the ApplicationMaster
4 - The ApplicationMaster asks the assigned NodeManagers to run the Map tasks
5 - The ApplicationMaster determines the number of Map tasks based in the Input Splits
6 - The ResourceManager Scheduler decides where to run the Map tasks based on memory requirements and data locality

Which of the following is the correct order in which these steps must run?

A. 2, 3, 5, 1, 6, 4
B. 1, 2, 3, 5, 4, 6
C. 6, 2, 1, 3, 4, 5
D. 2, 5, 1, 3, 6, 4

27. 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 directly 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 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.
D. The client contacts the NameNode for the block location(s). The NameNode then queries the DataNodes for the 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.

28. You need to create a GUI application to help your company's sales people add and edit customer information. Your plan is to maintain the customer information in a flat CSV file. Would HDFS be appropriate for this customer information file?

    A. Yes, because HDFS is optimized for random access writes
    B. Yes, because HDFS is optimized for fast retrieval of relatively small amounts of data
    C. No, because HDFS can only be accessed by MapReduce applications
    D. No, because HDFS is optimized for write-once, streaming access for relatively large files

29. You are configuring your cluster to run HDFS and MapReduce v2 (MRv2) on YARN. Which two daemons need to be installed on your cluster's master nodes?

    A. HMaster
    B. ResourceManager
    C. TaskManager
    D. JobTracker
    E. NameNode
    F. DataNode

30. Which YARN daemon or service negotiates map and reduce Containers from the, tracking their status and monitoring progress?

    A. NodeManager
    B. ApplicationMaster
    C. ApplicationManager
    D. ResourceManager

31. During the execution of a MapReduce v2 (MRv2) job on YARN, where does the Mapper place the intermediate data of each Map task?

    A. The Mapper stores the intermediate data on the node running the job's ApplicationMaster so that it is available to YARN ShuffleService before the data is presented to the Reducer
    B. The Mapper stores the intermediate data in HDFS on the node where the Map tasks ran in the HDFS /usercache/&(user)/apache/application_&(appid) directory for the user who ran the job
    C. The Mapper transfers the intermediate data immediately to the Reducers as it is generated by the Map task
    D. YARN holds the intermediate data in the NodeManager's memory (a container) until it is transferred to the Reducer
    E. The Mapper stores the intermediate data on the underlying filesystem of the local disk in the directories yarn.nodemanager.local-dirs

32. You have a cluster with 60GB of memory which is shared between three queues: Production, Marketing, and Engineering. Production is currently demanding 30 GB, Marketing wants 20GB, and Engineering wants 50GB. Given the configuration below, how will memory be allocated to each of the queues?

```
<allocations>
    <queue name="Production">
        <minResources>20000 mb, 0 vcores</minResources>
        <maxResources>50000 mb, 0 vcores</maxResources>
        <schedulingPolicy>fair</schedulingPolicy>
    </queue>
    <queue name="Marketing">
        <minResources>10000 mb, 0 vcores</minResources>
        <maxResources>40000 mb, 0 vcores</maxResources>
        <schedulingPolicy>fair</schedulingPolicy>
    </queue>
    <userMaxAppsDefault>5</userMaxAppsDefault>
</allocations>
```

    A. 10 GB for Marketing, 20 GB for Production, and 30 GB for Engineering
    B. 20 GB for Marketing, 20 GB for Production, and 20 GB for Engineering
    C. 10 GB for Marketing, 50 GB for Production, and 0 GB for Engineering
    D. 20 GB for Marketing, 40 GB for Production, and 0 GB for Engineering

33. 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?

    A. It only keeps track of which NameNode is Active at any given time
    B. It monitors an NFS mount point and reports if the mount point disappears
    C. 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
    D. It only manages the edits file which is a log of changes to the HDFS filesystem
    E. Clients connect to the ZooKeeper to determine which NameNode is Active

34. Choose three reasons why you should run the HDFS balancer periodically.

    A. To ensure that there is capacity in HDFS for additional data
    B. To ensure that all blocks in the cluster are 128 MB in size
    C. To help HDFS deliver consistent performance under heavy loads
    D. To ensure that there is consistent disk utilization across the DataNodes
    E. To improve data locality in MapReduce

35. Assuming default values, which of the following is the web user interface port for the NameNode?

    A. 19888
    B. 50075
    C. 50070
    D. 50090

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

    A. Install the impalad daemon, statestored daemon, and catalogd daemon on each machine in the cluster, and the impala shell on your gateway machine
    B. Install the impalad daemon, the statestored daemon, the catalogd daemon, and the impala shell on your gateway machine
    C. 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
    D. 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
    E. Install the impalad daemon, statestored daemon, and catalogd daemon on each machine in the cluster and on your gateway machine

37. In which of the following folders would you find the local logs for a NodeManager?

    A. /var/log/hadoop-hdfs/
    B. /var/log/hadoop-yarn/
    C. /var/log/hadoop-mapreduce/
    D. /var/log/impala

38. In which configuration file would you set the host name for the ResourceManager?

    A. core-site.xml
    B. hdfs-site.xml
    C. yarn-site.xml
    D. mapred-site.xml

39. What best describes the relationship between MapReduce and Hive?

    A. Hive provides no additional capabilities to MapReduce. Hive programs are executed as MapReduce jobs via the Hive interpreter.
    B. Hive provides the additional capability of allowing you to control the flow of multiple MapReduce jobs.
    C. Hive provides additional capabilities that allow certain types of data manipulation not possible with MapReduce
    D. Hive programs rely on MapReduce but are extensible, allowing developers to dos special-purpose processing not provided by MapReduce

40. You have some data in a folder in HDFS called /data/input. You want to use this data for use in Hive. How do you create a table in Hive to do this?

    A. hive> CREATE TABLE input LOCATION '/data/input';
    B. hive> CREATE EXTERNAL TABLE input LOCATION '/data/input';
    C. $ sqoop import --connect jdbc:hdfs://data/input --table input

END OF TEST