Cloudera

Exam CCA-410

Cloudera Certified Administrator for Apache Hadoop (CCAH)

Version: 7.5

[Total Questions: 97]

Question No : 1

Which two updates occur when a client application opens a stream to begin a file write on a cluster running MapReduce v1 (MRv1)?

A. Once the write stream closes on the DataNode, the DataNode immediately initiates a black report to the NameNode.
B. The change is written to the NameNode disk.
C. The metadata in the RAM on the NameNode is flushed to disk.
D. The metadata in RAM on the NameNode is flushed disk.
E. The metadata in RAM on the NameNode is updated.
F. The change is written to the edits file.

Answer: E,F

Question No : 2

For a MapReduce job, on a cluster running MapReduce v1 (MRv1), what's the relationship between tasks and task templates?

A. There are always at least as many task attempts as there are tasks.
B. There are always at most as many tasks attempts as there are tasks.
C. There are always exactly as many task attempts as there are tasks.
D. The developer sets the number of task attempts on job submission.

Answer: A

Question No : 3

What action occurs automatically on a cluster when a DataNode is marked as dead?

A. The NameNode forces re-replication of all the blocks which were stored on the dead DataNode.
B. The next time a client submits job that requires blocks from the dead DataNode, the JobTracker receives no heart beats from the DataNode. The JobTracker tells the NameNode that the DataNode is dead, which triggers block re-replication on the cluster.
C. The replication factor of the files which had blocks stored on the dead DataNode is temporarily reduced, until the dead DataNode is recovered and returned to the cluster.
D. The NameNode informs the client which write the blocks that are no longer available;

the client then re-writes the blocks to a different DataNode.

Answer: A

Explanation: How NameNode Handles data node failures?

NameNode periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster. Receipt of a Heartbeat implies that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode. When NameNode notices that it has not recieved a hearbeat message from a data node after a certain amount of time, the data node is marked as dead. Since blocks will be under replicated the system begins replicating the blocks that were stored on the dead datanode. The NameNode Orchestrates the replication of data blocks from one datanode to another. The replication data transfer happens directly between datanodes and the data never passes through the namenode.

Note: If the Name Node stops receiving heartbeats from a Data Node it presumes it to be dead and any data it had to be gone as well. Based on the block reports it had been receiving from the dead node, the Name Node knows which copies of blocks died along with the node and can make the decision to re-replicate those blocks to other Data Nodes. It will also consult the Rack Awareness data in order to maintain thetwo copies in one rack, one copy in another rackreplica rule when deciding which Data Node should receive a new copy of the blocks.

Reference: 24 Interview Questions & Answers for Hadoop MapReduce developers, How NameNode Handles data node failures'

Question No : 4

How does the NameNode know DataNodes are available on a cluster running MapReduce v1 (MRv1)

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

Answer: B

Explanation: How NameNode Handles data node failures?

NameNode periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster. Receipt of a Heartbeat implies that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode. When NameNode notices that it has not recieved a hearbeat message from a data node after a certain amount of time, the data node is marked as dead. Since blocks will be under replicated the system begins replicating the blocks that were stored on the dead datanode. The NameNode Orchestrates the replication of data blocks from one datanode to another. The replication data transfer happens directly between datanodes and the data never passes through the namenode.

Reference: 24 Interview Questions & Answers for Hadoop MapReduce developers, How NameNode Handles data node failures?

Question No : 5

Which three distcp features can you utilize on a Hadoop cluster?

A. Use distcp to copy files only between two clusters or more. You cannot use distcp to copy data between directories inside the same cluster.
B. Use distcp to copy HBase table files.
C. Use distcp to copy physical blocks from the source to the target destination in your cluster.
D. Use distcp to copy data between directories inside the same cluster.
E. Use distcp to run an internal MapReduce job to copy files.

Answer: B,D,E

Explanation:
DistCp (distributed copy) is a tool used for large inter/intra-cluster copying. It uses Map/Reduce to effect its distribution, error handling and recovery, and reporting. It expands a list of files and directories into input to map tasks, each of which will copy a partition of the files specified in the source list. Its Map/Reduce pedigree has endowed it with some quirks in both its semantics and execution.

Reference: Hadoop DistCp Guide

Question No : 6

How does HDFS Federation help HDFS Scale horizontally?

A. HDFS Federation improves the resiliency of HDFS in the face of network issues by removing the NameNode as a single-point-of-failure.
B. HDFS Federation allows the Standby NameNode to automatically resume the services of an active NameNode.
C. HDFS Federation provides cross-data center (non-local) support for HDFS, allowing a cluster administrator to split the Block Storage outside the local cluster.
D. HDFS Federation reduces the load on any single NameNode by using the multiple, independent NameNode to manage individual pars of the filesystem namespace.

Answer: D

Explanation: HDFS FederationIn order to scale the name service horizontally, federation uses multiple independent Namenodes/Namespaces. The Namenodes are federated, that is, the Namenodes are independent and don't require coordination with each other. The datanodes are used as common storage for blocks by all the Namenodes. Each datanode registers with all the Namenodes in the cluster. Datanodes send periodic heartbeats and block reports and handles commands from the Namenodes.

Reference: Apache Hadoop 2.0.2-alpha

http://hadoop.apache.org/docs/current/

Question No : 7

Choose which best describe a Hadoop cluster's block size storage parameters once you set the HDFS default block size to 64MB?

A. The block size of files in the cluster can be determined as the block is written.
B. The block size of files in the Cluster will all be multiples of 64MB.
C. The block size of files in the duster will all at least be 64MB.
D. The block size of files in the cluster will all be the exactly 64MB.

Answer: D

Explanation:

Note: What is HDFS Block size? How is it different from traditional file system block size?

In HDFS data is split into blocks and distributed across multiple nodes in the cluster. Each block is typically 64Mb or 128Mb in size. Each block is replicated multiple times. Default is to replicate each block three times. Replicas are stored on different nodes. HDFS utilizes the local file system to store each HDFS block as a separate file. HDFS Block size can not be compared with the traditional file system block size.

Question No : 8

Which MapReduce daemon instantiates user code, and executes map and reduce tasks on a cluster running MapReduce v1 (MRv1)?

A. NameNode
B. DataNode
C. JobTracker
D. TaskTracker
E. ResourceManager
F. ApplicationMaster
G. NodeManager

Answer: D

Explanation: A TaskTracker is a slave node daemon in the cluster that accepts tasks (Map, Reduce and Shuffle operations) from a JobTracker. There is only One Task Tracker process run on any hadoop slave node. Task Tracker runs on its own JVM process. Every TaskTracker is configured with a set of slots, these indicate the number of tasks that it can accept. The TaskTracker starts a separate JVM processes to do the actual work (called as Task Instance) this is to ensure that process failure does not take down the task tracker. The TaskTracker monitors these task instances, capturing the output and exit codes. When the Task instances finish, successfully or not, the task tracker notifies the JobTracker. The TaskTrackers also send out heartbeat messages to the JobTracker, usually every few minutes, to reassure the JobTracker that it is still alive. These message also inform the JobTracker of the number of available slots, so the JobTracker can stay up to date with where in the cluster work can be delegated.

Note: How many Daemon processes run on a Hadoop system?

Hadoop is comprised of five separate daemons. Each of these daemon run in its own JVM.
Following 3 Daemons run on Master
nodes NameNode - This daemon stores and maintains the metadata for HDFS.
Secondary NameNode - Performs housekeeping functions for the NameNode.
JobTracker - Manages MapReduce jobs, distributes individual tasks to machines running the Task Tracker.

Following 2 Daemons run on each Slave nodes
DataNode – Stores actual HDFS data blocks.
TaskTracker - Responsible for instantiating and monitoring individual Map and Reduce tasks.

Reference: 24 Interview Questions & Answers for Hadoop MapReduce developers, What is a Task Tracker in Hadoop? How many instances of TaskTracker run on a Hadoop Cluster

Question No : 9

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.

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 to the cluster.
C. You don't need to restart any daemon, as they will pick up changes automatically.
D. You must modify the configuration files on each of the six DataNode machines.
E. You must modify the configuration files on only one of the DataNode machine
F. You must modify the configuration files on the NameNode only. DataNodes read their configuration from the master nodes.

Answer: B,D

Question No : 10

Identify the function performed by the Secondary NameNode daemon on a cluster

configured to run with a single NameNode.

A. In this configuration, the Secondary NameNode performs a checkpoint operation on the files by the NameNode.
B. In this configuration, the Secondary NameNode is standby NameNode, ready to failover and provide high availability.
C. In this configuration, the Secondary NameNode performs deal-time backups of the NameNode.
D. In this configuration, the Secondary NameNode servers as alternate data channel for clients to reach HDFS, should the NameNode become too busy.

Answer: A

Explanation: The term "secondary name-node" is somewhat misleading. It is not a name-node in the sense that data-nodes cannot connect to the secondary name-node, and in no event it can replace the primary name-node in case of its failure.

The only purpose of the secondary name-node is to perform periodic checkpoints. The secondary name-node periodically downloads current name-node image and edits log files, joins them into new image and uploads the new image back to the (primary and the only) name-node.

So if the name-node fails and you can restart it on the same physical node then there is no need to shutdown data-nodes, just the name-node need to be restarted. If you cannot use the old node anymore you will need to copy the latest image somewhere else. The latest image can be found either on the node that used to be the primary before failure if available; or on the secondary name-node. The latter will be the latest checkpoint without subsequent edits logs, that is the most recent name space modifications may be missing there. You will also need to restart the whole cluster in this case.

Reference: Hadoop Wiki, What is the purpose of the secondary name-node?

Question No : 11

You install Cloudera Manager on a cluster where each host has 1 GB of RAM. All of the services show their status as concerning. However, all jobs submitted complete without an error.

Why is Cloudera Manager showing the concerning status KM the services?

A. A slave node's disk ran out of space
B. The slave nodes, haven't sent a heartbeat in 60 minutes
C. The slave nodes are swapping.
D. DataNode service instance has crashed.

Answer: B

Explanation: Concerning: There is an irregularity in the status of a service instance or role instance, but Cloudera Manager calculates that the instance might recover. For example, if the number of missed heartbeats exceeds a configurable threshold, the health status becomes Concerning. Or, if an instance is running on a host and the host is rebooted, the instance will be reported as In Progress for some period of time while it is restarting. Because the instance is expected to be Started, its health will be reported as Concerning until it transitions to started.

Note:
Bad: The service instance or role instance is not performing or did not finish performing the last command as expected, and Cloudera Manager calculates that the instance will not recover. For example, if the number of missed heartbeats exceeds a second (higher) configurable threshold, the health status becomes Bad. Another example of bad health is if a role you have stopped is actually still running, or a started role has stopped unexpectedly.

Good: The service instance or role instance is performing or has finished performing the last command as expected. This does not necessarily mean the service is running, it means it is behaving as expected. For example, if you clicked Stop to stop a role instance and it stopped successfully, then that role instance has a Good health status, even though it is not running.

Reference: About Service, Role, and Host Health

Question No : 12

What is the recommended disk configuration for slave nodes in your Hadoop cluster with 6 x 2 TB hard drives?

A. RAID 10
B. JBOD
C. RAID 5

D. RAID 1+0

Answer: B

Explanation:

Note: Let me be clear here…there are absolutely times when using a Enterprise-class storage device makes perfect sense. But for Hadoop it is very much unnecessary, and it is these three areas that I am going to hit as well as some others that I hope will demonstrate that Hadoop works best with inexpensive, internal storage in JBOD mode. Some of you might say “if you lose a disk in a JBOD configuration, you’re toast…you lose everything”. While this might be true, with Hadoop, it isn’t. Not only do you have the benefit that JBOD gives you in speed, you have the benefit that Hadoop Distributed File System (HDFS) negates this risk. HDFS basically creates three copies of the data. This is a very robust way to guard against data loss due to a disk failure or node outage, so you can eliminate the need for performance-reducing RAID.

Reference: Hadoop and Storage Area Networks

Question No : 13

You configure you cluster with HDFS High Availability (HA) using Quorum-Based storage. You do not implement HDFS Federation.

What is the maximum number of NameNodes daemon you should run on you cluster in order to avoid a “split-brain” scenario with your NameNodes?

A. Unlimited. HDFS High Availability (HA) is designed to overcome limitations on the number of NameNodes you can deploy.
B. Two active NameNodes and one Standby NameNode
C. One active NameNode and one Standby NameNode
D. Two active NameNodes and two Standby NameNodes

Answer: C

Explanation: In a typical HA cluster, two separate machines are configured as NameNodes. At any point in time, one of the NameNodes is in an Active state, and the other is in a Standby state. The Active NameNode is responsible for all client operations in the cluster, while the Standby is simply acting as a slave, maintaining enough state to provide a fast failover if necessary.