ISLEVER

CCA-470

Cloudera Certified Administrator for Apache Hadoop CDH4 Upgrade Exam (CCAH)

DEMO

https://www.islever.com/cca-470.html

https://www.islever.com/cloudera.html

For the most up-to-date exam questions and materials, we recommend visiting our website, where you can access the latest content and resources.

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: D,F
Explanation: Note: Namenode stores modifications to the filesystem as a log appended to a native filesystem file (edits). When a Namenode starts up, it reads HDFS state from an image file (fsimage) and then applies edits from edits log file. It then writes new HDFS state to (fsimage) and starts normal operation with an empty edits file. Since namenode merges fsimage and edits files only during start up, edits file could get very large over time on a large cluster. Another side effect of larger edits file is that next restart of Namenade takes longer.

The secondary namenode merges fsimage and edits log periodically and keeps edits log size with in a limit. It is usually run on a different machine than the primary Namenode since its memory requirements are on the same order as the primary namemode. The secondary namenode is started by bin/start-dfs.sh on the nodes specified in conf/masters file.

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: C
Explanation:

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 the two copies in one rack, one copy in another rack replica 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.