CCD-410^{Q&As}

Cloudera Certified Developer for Apache Hadoop (CCDH)

Pass Cloudera CCD-410 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.leads4pass.com/ccd-410.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Cloudera Official Exam Center

⚙ **Instant Download** After Purchase

⚙ **100% Money Back** Guarantee

⚙ **365 Days** Free Update

⚙ **800,000+** Satisfied Customers

QUESTION 1

You need to move a file titled "weblogs" into HDFS. When you try to copy the file, you can\\'t. You know you have ample space on your DataNodes. Which action should you take to relieve this situation and store more files in HDFS?

A. Increase the block size on all current files in HDFS.

B. Increase the block size on your remaining files.

C. Decrease the block size on your remaining files.

D. Increase the amount of memory for the NameNode.

E. Increase the number of disks (or size) for the NameNode.

F. Decrease the block size on all current files in HDFS.

Correct Answer: D

QUESTION 2

Indentify which best defines a SequenceFile?

A. A SequenceFile contains a binary encoding of an arbitrary number of homogeneous Writable objects

B. A SequenceFile contains a binary encoding of an arbitrary number of heterogeneous Writable objects

C. A SequenceFile contains a binary encoding of an arbitrary number of WritableComparable objects, in sorted order.

D. A SequenceFile contains a binary encoding of an arbitrary number key-value pairs. Each key must be the same type. Each value must be the same type.

Correct Answer: D

SequenceFile is a flat file consisting of binary key/value pairs.

There are 3 different SequenceFile formats:

Uncompressed key/value records.

Record compressed key/value records - only \\'values\\' are compressed here.

Block compressed key/value records - both keys and values are collected in \\'blocks\\' separately and

compressed. The size of the \\'block\\' is configurable.

Reference: http://wiki.apache.org/hadoop/SequenceFile

QUESTION 3

Identify the tool 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. Oozie

B. Flume

C. Pig

D. Hue

E. Hive

F. Sqoop

G. fuse-dfs

Correct Answer: F

Sqoop ("SQL-to-Hadoop") is a straightforward command-line tool with the following capabilities: Imports individual tables or entire databases to files in HDFS Generates Java classes to allow you to interact with your imported data Provides the ability to import from SQL databases straight into your Hive data warehouse Note: Data Movement Between Hadoop and Relational Databases Data can be moved between Hadoop and a relational database as a bulk data transfer, or relational tables

can be accessed from within a MapReduce map function. Note:

* Cloudera\\'s Distribution for Hadoop provides a bulk data transfer tool (i.e., Sqoop) that imports individual tables or entire databases into HDFS files. The tool also generates Java classes that support interaction with the imported data. Sqoop supports all relational databases over JDBC, and Quest Software provides a connector (i.e., OraOop) that has been optimized for access to data residing in Oracle databases.

Reference: http://log.medcl.net/item/2011/08/hadoop-and-mapreduce-big-data-analytics-gartner/ (Data Movement between hadoop and relational databases, second paragraph)

QUESTION 4

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 will be marked as corrupted if data node B fails during the creation of the file.

B. Each data node locks the local file to prohibit concurrent readers and writers of the file.

C. Each data node stores a copy of the file in the local file system with the same name as the HDFS file.

D. The file can be accessed if at least one of the data nodes storing the file is available.

Correct Answer: D

HDFS keeps three copies of a block on three different datanodes to protect against true data corruption.

HDFS also tries to distribute these three replicas on more than one rack to protect against data availability

issues. The fact that HDFS actively monitors any failed datanode(s) and upon failure detection immediately

schedules re-replication of blocks (if needed) implies that three copies of data on three different nodes is

sufficient to avoid corrupted files.

Note:

HDFS is designed to reliably store very large files across machines in a large cluster. It stores each file as

a sequence of blocks; all blocks in a file except the last block are the same size. The blocks of a file are

replicated for fault tolerance. The block size and replication factor are configurable per file. An application

can specify the number of replicas of a file. The replication factor can be specified at file creation time and

can be changed later. Files in HDFS are write-once and have strictly one writer at any time. The

NameNode makes all decisions regarding replication of blocks. HDFS uses rack-aware replica placement

policy. In default configuration there are total 3 copies of a datablock on HDFS, 2 copies are stored on

datanodes on same rack and 3rd copy on a different rack.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers , How the HDFS Blocks

are replicated?

QUESTION 5

How are keys and values presented and passed to the reducers during a standard sort and shuffle phase of
MapReduce?

A. Keys are presented to reducer in sorted order; values for a given key are not sorted.

B. Keys are presented to reducer in sorted order; values for a given key are sorted in ascending order.

C. Keys are presented to a reducer in random order; values for a given key are not sorted.

D. Keys are presented to a reducer in random order; values for a given key are sorted in ascending order.

Correct Answer: A

Reducer has 3 primary phases:

1.

 Shuffle

The Reducer copies the sorted output from each Mapper using HTTP across the network.

2.

 Sort

The framework merge sorts Reducer inputs by keys (since different Mappers may have output the same key).

The shuffle and sort phases occur simultaneously i.e. while outputs are being fetched they are merged.

SecondarySort

To achieve a secondary sort on the values returned by the value iterator, the application should extend the key with the secondary key and define a grouping comparator. The keys will be sorted using the entire key, but will be grouped using the grouping comparator to decide which keys and values are sent in the same call to reduce.

3. Reduce

In this phase the reduce(Object, Iterable, Context) method is called for each

in the sorted inputs.

The output of the reduce task is typically written to a RecordWriter via TaskInputOutputContext.write

(Object, Object).

The output of the Reducer is not re-sorted.

Reference: org.apache.hadoop.mapreduce, Class

Reducer

QUESTION 6

Your cluster\\'s HDFS block size in 64MB. You have directory containing 100 plain text files, each of which is 100MB in size. The InputFormat for your job is TextInputFormat. Determine how many Mappers will run?

A. 64

B. 100

C. 200

D. 640

Correct Answer: C

Each file would be split into two as the block size (64 MB) is less than the file size (100 MB), so 200 mappers would be running.

Note:

If you\\'re not compressing the files then hadoop will process your large files (say 10G), with a number of mappers related to the block size of the file.

Say your block size is 64M, then you will have ~160 mappers processing this 10G file (160*64 ~= 10G). Depending on how CPU intensive your mapper logic is, this might be an acceptable blocks size, but if you find that your mappers are executing in sub minute times, then you might want to increase the work done by each mapper (by increasing the block size to 128, 256, 512m - the actual size depends on how you intend to process the data).

Reference: http://stackoverflow.com/questions/11014493/hadoop-mapreduce-appropriate-input-files-size (first answer, second paragraph)

QUESTION 7

In a large MapReduce job with m mappers and n reducers, how many distinct copy operations will there be in the sort/shuffle phase?

A. mXn (i.e., m multiplied by n)

B. n

C. m

D. m+n (i.e., m plus n)

E. mn (i.e., m to the power of n)

Correct Answer: A

A MapReduce job with m mappers and r reducers involves up to m * r distinct copy operations, since each mapper may have intermediate output going to every reducer.

QUESTION 8

What data does a Reducer reduce method process?

A. All the data in a single input file.

B. All data produced by a single mapper.

C. All data for a given key, regardless of which mapper(s) produced it.

D. All data for a given value, regardless of which mapper(s) produced it.

Correct Answer: C

Reducing lets you aggregate values together. A reducer function receives an iterator of input values from

an input list. It then combines these values together, returning a single output value.

All values with the same key are presented to a single reduce task.

Reference: Yahoo! Hadoop Tutorial, Module 4: MapReduce

QUESTION 9

Indentify the utility that allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer?

A. Oozie

B. Sqoop

C. Flume

D. Hadoop Streaming

E. mapred

Correct Answer: D

Hadoop streaming is a utility that comes with the Hadoop distribution. The utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

Reference: http://hadoop.apache.org/common/docs/r0.20.1/streaming.html (Hadoop Streaming, second sentence)

QUESTION 10

What is the disadvantage of using multiple reducers with the default HashPartitioner and distributing your workload across you cluster?

A. You will not be able to compress the intermediate data.

B. You will longer be able to take advantage of a Combiner.

C. By using multiple reducers with the default HashPartitioner, output files may not be in globally sorted order.

D. There are no concerns with this approach. It is always advisable to use multiple reduces.

Correct Answer: C

Multiple reducers and total ordering

If your sort job runs with multiple reducers (either because mapreduce.job.reduces in mapred- site.xml has been set to a number larger than 1, or because you\\\'ve used the -r option to specify the number of reducers on the command-line), then by default Hadoop will use the HashPartitioner to distribute records across the reducers. Use of the HashPartitioner means that you can\\\'t concatenate your output files to create a single sorted output file. To do this you\\\'ll need total ordering,

Reference: Sorting text files with MapReduce

[CCD-410 VCE Dumps](#)   [CCD-410 Study Guide](#)   [CCD-410 Exam Questions](#)