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.pass4itsure.com/ccd-410.html

100% Passing Guarantee
100% Money Back Assurance

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

QUESTION 1

You need to create a job that does frequency analysis on input data. You will do this by writing a Mapper that uses TextInputFormat and splits each value (a line of text from an input file) into individual characters. For each one of these characters, you will emit the character as a key and an InputWritable as the value. As this will produce proportionally more intermediate data than input data, which two resources should you expect to be bottlenecks?

A. Processor and network I/O

B. Disk I/O and network I/O

C. Processor and RAM

D. Processor and disk I/O

Correct Answer: B

QUESTION 2

When is the earliest point at which the reduce method of a given Reducer can be called?

A. As soon as at least one mapper has finished processing its input split.

B. As soon as a mapper has emitted at least one record.

C. Not until all mappers have finished processing all records.

D. It depends on the InputFormat used for the job.

Correct Answer: C

In a MapReduce job reducers do not start executing the reduce method until the all Map jobs have completed. Reducers start copying intermediate key-value pairs from the mappers as soon as they are available. The programmer defined reduce method is called only after all the mappers have finished.

Note: The reduce phase has 3 steps: shuffle, sort, reduce. Shuffle is where the data is collected by the reducer from each mapper. This can happen while mappers are generating data since it is only a data transfer. On the other hand, sort and reduce can only start once all the mappers are done.

Why is starting the reducers early a good thing? Because it spreads out the data transfer from the mappers to the reducers over time, which is a good thing if your network is the bottleneck.

Why is starting the reducers early a bad thing? Because they "hog up" reduce slots while only copying data. Another job that starts later that will actually use the reduce slots now can\\'t use them.

You can customize when the reducers startup by changing the default value of mapred.reduce.slowstart.completed.maps in mapred-site.xml. A value of 1.00 will wait for all the mappers to finish before starting the reducers. A value of 0.0 will start the reducers right away. A value of 0.5 will start the reducers when half of the mappers are complete. You can also change mapred.reduce.slowstart.completed.maps on a job-by-job basis. Typically, keep mapred.reduce.slowstart.completed.maps above 0.9 if the system ever has multiple jobs running at once. This way the job doesn\\'t hog up reducers when they aren\\'t doing anything but copying data. If you only ever

have one job running at a time, doing 0.1 would probably be appropriate.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers, When is the reducers are started in a MapReduce job?

QUESTION 3

In a MapReduce job, the reducer receives all values associated with same key. Which statement best describes the ordering of these values?

A. The values are in sorted order.

B. The values are arbitrarily ordered, and the ordering may vary from run to run of the same MapReduce job.

C. The values are arbitrary ordered, but multiple runs of the same MapReduce job will always have the same ordering.

D. Since the values come from mapper outputs, the reducers will receive contiguous sections of sorted values.

Correct Answer: B

Note:

*

Input to the Reducer is the sorted output of the mappers.

*

The framework calls the application\\'s Reduce function once for each unique key in the sorted order.

*

Example:

For the given sample input the first map emits:

The second map emits:

QUESTION 4

In the reducer, the MapReduce API provides you with an iterator over Writable values. What does calling the next () method return?

A. It returns a reference to a different Writable object time.

B. It returns a reference to a Writable object from an object pool.

C. It returns a reference to the same Writable object each time, but populated with different data.

D. It returns a reference to a Writable object. The API leaves unspecified whether this is a reused object or a new object.

E. It returns a reference to the same Writable object if the next value is the same as the previous value, or a new Writable object otherwise.

Correct Answer: C

Calling Iterator.next() will always return the SAME EXACT instance of IntWritable, with the contents of that instance replaced with the next value.

Reference: manupulating iterator in mapreduce

QUESTION 5

You want to populate an associative array in order to perform a map-side join. You\\'ve decided to put this information in a text file, place that file into the DistributedCache and read it in your Mapper before any records are processed.

Indentify which method in the Mapper you should use to implement code for reading the file and populating the associative array?

A. combine

B. map

C. init

D. configure

Correct Answer: B

Reference: org.apache.hadoop.filecache , Class DistributedCache

QUESTION 6

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 7

You have user profile records in your OLPT database, that you want to join with web logs you have already ingested into the Hadoop file system. How will you obtain these user records?

A. HDFS command

B. Pig LOAD command

C. Sqoop import

D. Hive LOAD DATA command

E. Ingest with Flume agents F. Ingest with Hadoop Streaming

Correct Answer: C

Reference: Hadoop and Pig for Large-Scale Web Log Analysis

QUESTION 8

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 9

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 10

You want to perform analysis on a large collection of images. You want to store this data in HDFS and process it with MapReduce but you also want to give your data analysts and data scientists the ability to process the data directly from HDFS with an interpreted high-level programming language like Python. Which format should you use to store this data in HDFS?

A. SequenceFiles

B. Avro

C. JSON

D. HTML

E. XML

F. CSV

Correct Answer: B

Reference: Hadoop binary files processing introduced by image duplicates finder