

100% Money Back
Guarantee

Vendor:Cloudera

Exam Code:CCD-410

Exam Name:Cloudera Certified Developer for Apache Hadoop (CCDH)

Version:Demo

QUESTION 1

When can a reduce class also serve as a combiner without affecting the output of a MapReduce program?

- A. When the types of the reduce operation's input key and input value match the types of the reducer's output key and output value and when the reduce operation is both communicative and associative.
- B. When the signature of the reduce method matches the signature of the combine method.
- C. Always. Code can be reused in Java since it is a polymorphic object-oriented programming language.
- D. Always. The point of a combiner is to serve as a mini-reducer directly after the map phase to increase performance.
- E. Never. Combiners and reducers must be implemented separately because they serve different purposes.

Correct Answer: A

You can use your reducer code as a combiner if the operation performed is commutative and associative.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers, What are combiners? When should I use a combiner in my MapReduce Job?

QUESTION 2

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 3

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.

Identify 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 4

Your cluster's HDFS block size is 64MB. You have a 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 ($160 \times 64 \approx 10G$). Depending on how CPU intensive your mapper logic is, this might be an acceptable block 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 5

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 directory 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 then queries the DataNodes for 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.
- D. 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.

Correct Answer: A

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers, How the Client communicates with HDFS?

QUESTION 6

You are developing a combiner that takes as input Text keys, IntWritable values, and emits Text keys, IntWritable values. Which interface should your class implement?

- A. Combiner
- B. Mapper
- C. Reducer
- D. Reducer
- E. Combiner

Correct Answer: D

QUESTION 7

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 8

What types of algorithms are difficult to express in MapReduce v1 (MRv1)?

- A. Algorithms that require applying the same mathematical function to large numbers of individual binary records.
- B. Relational operations on large amounts of structured and semi-structured data.
- C. Algorithms that require global, sharing states.
- D. Large-scale graph algorithms that require one-step link traversal.
- E. Text analysis algorithms on large collections of unstructured text (e.g, Web crawls).

Correct Answer: C

See 3) below.

Limitations of Mapreduce where not to use Mapreduce While very powerful and applicable to a wide variety of problems, MapReduce is not the answer to every problem. Here are some problems I found where MapReudce is not suited and some papers that address the limitations of MapReuce.

1.

Computation depends on previously computed values

If the computation of a value depends on previously computed values, then MapReduce cannot be used. One good example is the Fibonacci series where each value is summation of the previous two values. i.e., $f(k+2) = f(k+1) + f(k)$. Also, if the data set is small enough to be computed on a single machine, then it is better to do it as a single `reduce(map(data))` operation rather than going through the entire map reduce process.

2.

Full-text indexing or ad hoc searching

The index generated in the Map step is one dimensional, and the Reduce step must not generate a large amount of data or there will be a serious performance degradation. For example, CouchDB's MapReduce may not be a good fit for full-text indexing or ad hoc searching. This is a problem better suited for a tool such as Lucene.

3.

Algorithms depend on shared global state

Solutions to many interesting problems in text processing do not require global synchronization. As a result, they can be expressed naturally in MapReduce, since map and reduce tasks run independently and in isolation. However, there are many examples of algorithms that depend crucially on the existence of shared global state during processing, making them difficult to implement in MapReduce (since the single opportunity for global synchronization in MapReduce is the barrier between the map and reduce phases of processing)

Reference: Limitations of Mapreduce where not to use Mapreduce

QUESTION 9

Given a directory of files with the following structure: line number, tab character, string:

Example: 1 abialkijfkaoasdfjksdlkjhqweroij 2 kadfjhuwqounahagtnbvaswslmnbfgy 3 kjfteiomndscxeqalkzhtopedkfsikj

You want to send each line as one record to your Mapper. Which InputFormat should you use to complete the line:
`conf.setInputFormat (____.class) ; ?`

- A. SequenceFileAsTextInputFormat
- B. SequenceFileInputFormat
- C. KeyValueFileInputFormat
- D. BDBInputFormat

Correct Answer: C

<http://stackoverflow.com/questions/9721754/how-to-parse-customwritable-from-text-in-hadoop>

QUESTION 10

You need to run the same job many times with minor variations. Rather than hardcoding all job configuration options in

your driver code, you've decided to have your Driver subclass `org.apache.hadoop.conf.Configuration` and implement the `org.apache.hadoop.util.Tool` interface. Identify which invocation correctly passes `mapred.job.name` with a value of Example to Hadoop?

- A. `hadoop "mapred.job.name=Example" MyDriver` input output
- B. `hadoop MyDriver mapred.job.name=Example` input output
- C. `hadoop MyDriver -D mapred.job.name=Example` input output
- D. `hadoop setproperty mapred.job.name=Example MyDriver` input output
- E. `hadoop setproperty ("mapred.job.name=Example") MyDriver` input output

Correct Answer: C

Configure the property using the `-D key=value` notation:

```
-D mapred.job.name='My Job'
```

You can list a whole bunch of options by calling the streaming jar with just the `-info` argument

Reference: [Python hadoop streaming : Setting a job name](#)

QUESTION 11

All keys used for intermediate output from mappers must:

- A. Implement a splittable compression algorithm.
- B. Be a subclass of `FileInputFormat`.
- C. Implement `WritableComparable`.
- D. Override `isSplittable`.
- E. Implement a comparator for speedy sorting.

Correct Answer: C

The MapReduce framework operates exclusively on pairs, that is, the framework views the input to the job as a set of pairs and produces a set of pairs as the output of the job, conceivably of different types.

The key and value classes have to be serializable by the framework and hence need to implement the `Writable` interface. Additionally, the key classes have to implement the `WritableComparable` interface to facilitate sorting by the framework.

Reference: [MapReduce Tutorial](#)

QUESTION 12

To process input key-value pairs, your mapper needs to load a 512 MB data file in memory. What is the best way to accomplish this?

- A. Serialize the data file, insert in it the JobConf object, and read the data into memory in the configure method of the mapper.
- B. Place the data file in the DistributedCache and read the data into memory in the map method of the mapper.
- C. Place the data file in the DataCache and read the data into memory in the configure method of the mapper.
- D. Place the data file in the DistributedCache and read the data into memory in the configure method of the mapper.

Correct Answer: D