



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.lead4pass.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

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: manipulating iterator in mapreduce

QUESTION 2

MapReduce v2 (MRv2/YARN) is designed to address which two issues?

- A. Single point of failure in the NameNode.
- B. Resource pressure on the JobTracker.
- C. HDFS latency.
- D. Ability to run frameworks other than MapReduce, such as MPI.
- E. Reduce complexity of the MapReduce APIs.
- F. Standardize on a single MapReduce API.

Correct Answer: BD

YARN (Yet Another Resource Negotiator), as an aspect of Hadoop, has two major kinds of benefits:

*

(D) The ability to use programming frameworks other than MapReduce. / MPI (Message Passing Interface) was mentioned as a paradigmatic example of a MapReduce alternative

*

Scalability, no matter what programming framework you use. Note:

*



The fundamental idea of MRv2 is to split up the two major functionalities of the JobTracker, resource management and job scheduling/monitoring, into separate daemons. The idea is to have a global ResourceManager (RM) and per-application ApplicationMaster (AM). An application is either a single job in the classical sense of Map-Reduce jobs or a DAG of jobs.

*

(B) The central goal of YARN is to clearly separate two things that are unfortunately smushed together in current Hadoop, specifically in (mainly) JobTracker:

/ Monitoring the status of the cluster with respect to which nodes have which resources available.

Under YARN, this will be global.

/ Managing the parallelization execution of any specific job. Under YARN, this will be done separately for each job. The current Hadoop MapReduce system is fairly scalable -- Yahoo runs 5000 Hadoop jobs, truly concurrently, on a single

cluster, for a total 1.5 2 millions jobs/ cluster/month. Still, YARN will remove scalability bottlenecks

Reference: Apache Hadoop YARN Concepts and Applications

QUESTION 3

Assuming default settings, which best describes the order of data provided to a reducer's reduce method:

- A. The keys given to a reducer aren't in a predictable order, but the values associated with those keys always are.
- B. Both the keys and values passed to a reducer always appear in sorted order.
- C. Neither keys nor values are in any predictable order.
- D. The keys given to a reducer are in sorted order but the values associated with each key are in no predictable order

Correct Answer: D

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 4

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 5

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

QUESTION 6

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 MapReduce is not suited and some papers that address the limitations of MapReduce.

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 7

You are developing a MapReduce job for sales reporting. The mapper will process input keys representing the year (IntWritable) and input values representing product identifiers (Text). Identify what determines the data types used by the Mapper for a given job.

- A. The key and value types specified in the JobConf.setMapInputKeyClass and JobConf.setMapInputValuesClass methods
- B. The data types specified in HADOOP_MAP_DATATYPES environment variable
- C. The mapper-specification.xml file submitted with the job determine the mapper's input key and value types.
- D. The InputFormat used by the job determines the mapper's input key and value types.

Correct Answer: D

The input types fed to the mapper are controlled by the InputFormat used. The default input format, "TextInputFormat," will load data in as (LongWritable, Text) pairs. The long value is the byte offset of the line in the file. The Text object holds



the string contents of the line of the file.

Note: The data types emitted by the reducer are identified by `setOutputKeyClass()` and `setOutputValueClass()`. The data types emitted by the reducer are identified by `setOutputKeyClass()` and `setOutputValueClass()`.

By default, it is assumed that these are the output types of the mapper as well. If this is not the case, the methods `setMapOutputKeyClass()` and `setMapOutputValueClass()` methods of the `JobConf` class will override these.

Reference: Yahoo! Hadoop Tutorial, THE DRIVER METHOD

QUESTION 8

MapReduce v2 (MRv2/YARN) splits which major functions of the JobTracker into separate daemons? Select two.

- A. Health states checks (heartbeats)
- B. Resource management
- C. Job scheduling/monitoring
- D. Job coordination between the ResourceManager and NodeManager
- E. Launching tasks
- F. Managing file system metadata
- G. MapReduce metric reporting
- H. Managing tasks

Correct Answer: BC

The fundamental idea of MRv2 is to split up the two major functionalities of the JobTracker, resource management and job scheduling/monitoring, into separate daemons. The idea is to have a global ResourceManager (RM) and per-application ApplicationMaster (AM). An application is either a single job in the classical sense of Map-Reduce jobs or a DAG of jobs.

Note:

The central goal of YARN is to clearly separate two things that are unfortunately smushed together in current Hadoop, specifically in (mainly) JobTracker:

/ Monitoring the status of the cluster with respect to which nodes have which resources available.

Under YARN, this will be global.

/ Managing the parallelization execution of any specific job. Under YARN, this will be done separately for each job.

Reference: Apache Hadoop YARN Concepts and Applications

QUESTION 9



Can you use MapReduce to perform a relational join on two large tables sharing a key? Assume that the two tables are formatted as comma-separated files in HDFS.

- A. Yes.
- B. Yes, but only if one of the tables fits into memory
- C. Yes, so long as both tables fit into memory.
- D. No, MapReduce cannot perform relational operations.
- E. No, but it can be done with either Pig or Hive.

Correct Answer: A

Note:

*

Join Algorithms in MapReduce

- A) Reduce-side join
- B) Map-side join
- C) In-memory join

/ Striped Striped variant variant

/ Memcached variant

*

Which join to use?

/ In-memory join > map-side join > reduce-side join

/ Limitations of each?

In-memory join: memory

Map-side join: sort order and partitioning

Reduce-side join: general purpose

QUESTION 10

You wrote a map function that throws a runtime exception when it encounters a control character in input data. The input supplied to your mapper contains twelve such characters total, spread across five file splits. The first four file splits each have two control characters and the last split has four control characters.

Identify the number of failed task attempts you can expect when you run the job with `mapred.max.map.attempts` set to 4:

- A. You will have forty-eight failed task attempts



- B. You will have seventeen failed task attempts
- C. You will have five failed task attempts
- D. You will have twelve failed task attempts
- E. You will have twenty failed task attempts

Correct Answer: E

There will be four failed task attempts for each of the five file splits. Note:

When the jobtracker is notified of a task attempt that has failed (by the tasktracker's heartbeat call), it will reschedule execution of the task. The jobtracker will try to avoid rescheduling the task on a tasktracker where it has previously failed. Furthermore, if a task fails four times (or more), it will not be retried further. This value is configurable: the maximum number of attempts to run a task is controlled by the `mapred.map.max.attempts` property for map tasks and `mapred.reduce.max.attempts` for reduce tasks. By default, if any task fails four times (or whatever the maximum number of attempts is configured to), the whole job fails.



QUESTION 11

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 12

Identify 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 13

You have the following key-value pairs as output from your Map task:

(the, 1) (fox, 1) (faster, 1) (than, 1) (the, 1) (dog, 1)

How many keys will be passed to the Reducer's reduce method?

- A. Six
- B. Five
- C. Four
- D. Two
- E. One



F. Three

Correct Answer: B

Only one key value pair will be passed from the two (the, 1) key value pairs.

QUESTION 14

You use the `hadoop fs put` command to write a 300 MB file using and HDFS block size of 64 MB. Just after this command has finished writing 200 MB of this file, what would another user see when trying to access this file?

- A. They would see Hadoop throw a `ConcurrentFileAccessException` when they try to access this file.
- B. They would see the current state of the file, up to the last bit written by the command.
- C. They would see the current of the file through the last completed block.
- D. They would see no content until the whole file written and closed.

Correct Answer: C

QUESTION 15

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.Configured` 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



To Read the [Whole Q&As](#), please purchase the [Complete Version](#) from [Our website](#).

Try our product !

100% Guaranteed Success
100% Money Back Guarantee
365 Days Free Update
Instant Download After Purchase
24x7 Customer Support
Average 99.9% Success Rate
More than 800,000 Satisfied Customers Worldwide
Multi-Platform capabilities - [Windows](#), [Mac](#), [Android](#), [iPhone](#), [iPod](#), [iPad](#), [Kindle](#)

We provide exam PDF and VCE of Cisco, Microsoft, IBM, CompTIA, Oracle and other IT Certifications. You can view Vendor list of All Certification Exams offered:

<https://www.lead4pass.com/allproducts>

Need Help

Please provide as much detail as possible so we can best assist you.
To update a previously submitted ticket:



 <p>One Year Free Update Free update is available within One Year after your purchase. After One Year, you will get 50% discounts for updating. And we are proud to boast a 24/7 efficient Customer Support system via Email.</p>	 <p>Money Back Guarantee To ensure that you are spending on quality products, we provide 100% money back guarantee for 30 days from the date of purchase.</p>	 <p>Security & Privacy We respect customer privacy. We use McAfee's security service to provide you with utmost security for your personal information & peace of mind.</p>
---	---	--

Any charges made through this site will appear as Global Simulators Limited.
All trademarks are the property of their respective owners.
Copyright © lead4pass, All Rights Reserved.