Free Questions for HDPCD by dumpshq

Shared by Greene on 29-01-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

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

Example:

1 abialkjfjkaoasdfjksdlkjhqweroij

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) ; ?

Options:

A- SequenceFileAsTextInputFormat

B- SequenceFileInputFormat

C- KeyValueFileInputFormat

D- BDBInputFormat

Answer:

C

Explanation:

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

Question 2

Question Type: MultipleChoice

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

Options:

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.

Answer:

C

Explanation:

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,

Question 3

Question Type: **MultipleChoice**

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?

Options:

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

Answer:

C

Question 4

Question Type: **MultipleChoice**

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?

Options:

A- Six

B- Five

C- Four

D- Two

E- One

F- Three

Answer:

B

Explanation:

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

Question 5

Question Type: MultipleChoice

For each input key-value pair, mappers can emit:

Options:

A- As many intermediate key-value pairs as designed. There are no restrictions on the types of those key-value pairs (i.e., they can be heterogeneous).

B- As many intermediate key-value pairs as designed, but they cannot be of the same type as the input key-value pair.

C- One intermediate key-value pair, of a different type.

D- One intermediate key-value pair, but of the same type.

E- As many intermediate key-value pairs as designed, as long as all the keys have the same types and all the values have the same type.

Answer:

E

Explanation:

Mapper maps input key/value pairs to a set of intermediate key/value pairs.

Maps are the individual tasks that transform input records into intermediate records. The transformed intermediate records do not need to be of the same type as the input records. A given input pair may map to zero or many output pairs.

Question 6

Question Type: MultipleChoice

Which best describes how TextInputFormat processes input files and line breaks?

Options:

A- Input file splits may cross line breaks. A line that crosses file splits is read by the RecordReader of the split that contains the

beginning of the broken line.

B- Input file splits may cross line breaks. A line that crosses file splits is read by the RecordReaders of both splits containing the broken line.

C- The input file is split exactly at the line breaks, so each RecordReader will read a series of complete lines.

D- Input file splits may cross line breaks. A line that crosses file splits is ignored.

E- Input file splits may cross line breaks. A line that crosses file splits is read by the RecordReader of the split that contains the end of the broken line.

Answer:

A

Question 7

Question Type: MultipleChoice

Identify the MapReduce v2 (MRv2 / YARN) daemon responsible for launching application containers and monitoring application resource usage?

Options:

A- ResourceManager

B- NodeManager

C- ApplicationMaster

D- ApplicationMasterService

E- TaskTracker

F- JobTracker

Answer:

B

Question 8

Question Type: **MultipleChoice**

You are developing a MapReduce job for sales reporting. The mapper will process input keys representing the year (IntWritable) and input values representing product indentifies (Text).

Indentify what determines the data types used by the Mapper for a given job.

Options:

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.

Answer:

D

Explanation:

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() andsetOutputValueClass(). 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.

Question 9

Question Type: MultipleChoice

You need to run the same job many times with minor variations. Rather than hardcoding all job configuration options in your drive code, you've decided to have your Driver subclass org.apache.hadoop.conf.Configured and implement the org.apache.hadoop.util.Tool interface.

Indentify which invocation correctly passes.mapred.job.name with a value of Example to Hadoop?

Options:

A- hadoop "mapred.job.name=Example" MyDriver input output

B- hadoop MyDriver mapred.job.name=Example input output

C- hadoop MyDrive --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

Answer:

C

Explanation:

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

Question 10

Question Type: MultipleChoice

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

Options:

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.

Answer:

A, B

Question 11

Question Type: **MultipleChoice**

You want to understand more about how users browse your public website, such as which pages they visit prior to placing an order. You have a farm of 200 web servers hosting your website. How will you gather this data for your analysis?

Options:

A- Ingest the server web logs into HDFS using Flume.

B- Write a MapReduce job, with the web servers for mappers, and the Hadoop cluster nodes for reduces.

C- Import all users' clicks from your OLTP databases into Hadoop, using Sqoop.

D- Channel these clickstreams inot Hadoop using Hadoop Streaming.

E- Sample the weblogs from the web servers, copying them into Hadoop using curl.

Answer:

A

Question 12

Question Type: MultipleChoice

You have just executed a MapReduce job. Where is intermediate data written to after being emitted from the Mapper's map method?

Options:

A- Intermediate data in streamed across the network from Mapper to the Reduce and is never written to disk.

B- Into in-memory buffers on the TaskTracker node running the Mapper that spill over and are written into HDFS.

C- Into in-memory buffers that spill over to the local file system of the TaskTracker node running the Mapper.

D- Into in-memory buffers that spill over to the local file system (outside HDFS) of the TaskTracker node running the Reducer

E- Into in-memory buffers on the TaskTracker node running the Reducer that spill over and are written into HDFS.

Answer:

C

Explanation:

The mapper output (intermediate data) is stored on the Local file system (NOT HDFS) of each individual mapper nodes. This is typically a temporary directory location which can be setup in config by the hadoop administrator. The intermediate data is cleaned up after the Hadoop Job completes.