Cloudera

CCD-410

Cloudera Certified Developer for Apache Hadoop (CCDH)

Class Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT>

QUESTION: 56
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.

Answer: A

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

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

QUESTION: 57
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

Answer: A

Explanation:
Using Hadoop Sequence Files
So what should we do in order to deal with huge amount of images? Use hadoop sequence files! Those are map files that inherently can be read by map reduce applications – there is an input format especially for sequence files – and are splitable by map reduce, so we can have one huge file that will be the input of many map tasks. By using those sequence files we are letting hadoop use its advantages. It can split the work into chunks so the processing is parallel, but the chunks are big enough that the process stays efficient. Since the sequence file are map file the desired format will be that the key will be text and hold the HDFS filename and the value will be BytesWritable and will contain the image content of the file.

Reference:
Hadoop binary files processing introduced by image duplicates finder

QUESTION: 58
You want to run Hadoop jobs on your development workstation for testing before you submit them to your production cluster. Which mode of operation in Hadoop allows you to most closely simulate a production cluster while using a single machine?

A. Run all the nodes in your production cluster as virtual machines on your development workstation.
B. Run the hadoop command with the –jt local and the –fs file:///options.
C. Run the DataNode, TaskTracker, NameNode and JobTracker daemons on a single machine.
D. Run simldooop, the Apache open-source software for simulating Hadoop clusters.

Answer: A

Explanation:
Hosting on local VMs
As well as large-scale cloud infrastructures, there is another deployment pattern: local VMs on desktop systems or other development machines. This is a good tactic if your physical machines run windows and you need to bring up a Linux system running Hadoop, and/or you want to simulate the complexity of a small Hadoop cluster.
Have enough RAM for the VM to not swap.
Don't try and run more than one VM per physical host, it will only make things slower.
use file: URLs to access persistent input and output data.
consider making the default filesystem a file: URL so that all storage is really on the physical host. It's often faster and preserves data better.

QUESTION: 59

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

Answer: C

Explanation:
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: 60
What is 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 same type.

Answer: D

Explanation:
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