ISLEVER

CCD-333

Cloudera Certified Developer for Apache Hadoop

DEMO

https://www.islever.com/ccd-333.html

https://www.islever.com/cloudera.html

For the most up-to-date exam questions and materials, we recommend visiting our website, where you can access the latest content and resources.

QUESTION NO: 1

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

QUESTION NO: 2

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

Example:

1. abialkjfjkaoasdfjksdlkjhqweroij

2. kadf jhuwqounahagtnbvaswslmnbfgy

3. kjfteiomndscxeqalkzhtopedkfslkj

You want to send each line as one record to your Mapper. Which InputFormat would you use to complete the line: setInputFormat (_____.class);

A. BDBInputFormat
B. KeyValueTextInputFormat

C. SequenceFileInputFormat
D. SequenceFileAsTextInputFormat

Answer: C

Explanation: Note:
The output format for your first MR job should be SequenceFileOutputFormat - this will store the Key/Values output from the reducer in a binary format, that can then be read back in, in your second MR job using SequenceFileInputFormat.

Reference:http://stackoverflow.com/questions/9721754/how-to-parse-customwritable-from-text-in-hadoop(see answer 1 and then see the comment #1 for it)

QUESTION NO: 3

In a MapReduce job, you want each of you input files processed by a single map task. How do you configure a MapReduce job so that a single map task processes each input file regardless of how many blocks the input file occupies?

A. Increase the parameter that controls minimum split size in the job configuration.
B. Write a custom MapRunner that iterates over all key-value pairs in the entire file.
C. Set the number of mappers equal to the number of input files you want to process.
D. Write a custom FileInputFormat and override the method isSplittable to always return false.

Answer: D

Explanation: Note:
*// Do not allow splitting.
protected boolean isSplittable(JobContext context, Path filename) {
return false;
}
*InputSplits: An InputSplit describes a unit of work that comprises a single map task in a MapReduce program. A MapReduce program applied to a data set, collectively referred to as a Job, is made up of several (possibly several hundred) tasks. Map tasks may involve reading a whole file; they often involve reading only part of a file. By default, the FileInputFormat and its descendants break a file up into 64 MB chunks (the same size as blocks in HDFS). You can control this value by setting the mapred.min.split.size parameter in hadoop-site.xml, or by overriding the parameter in the JobConf object used to submit a particular MapReduce job. By processing a file in chunks, we allow several map tasks to operate on a single file in parallel. If the file is very large, this can improve performance significantly through parallelism. Even more importantly, since the various blocks that make up the file may be spread across several different nodes in the cluster, it allows tasks to be scheduled on each of these different nodes; the

individual blocks are thus all processed locally, instead of needing to be transferred from one node to another. Of course, while log files can be processed in this piece-wise fashion, some file formats are not amenable to chunked processing. By writing a custom InputFormat, you can control how the file is broken up (or is not broken up) into splits.

QUESTION NO: 4

Which of the following best describes the workings of TextInputFormat?

A. Input file splits may cross line breaks. A line that crosses tile splits is ignored.
B. The input file is split exactly at the line breaks, so each Record Reader will read a series of complete lines.
C. 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.
D. 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.
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 beginning of the broken line.

Answer: D
Explanation: As the Map operation is parallelized the input file set is first split to several pieces called FileSplits. If an individual file is so large that it will affect seek time it will be split to several Splits. The splitting does not know anything about the input file's internal logical structure, for example line-oriented text files are split on arbitrary byte boundaries. Then a new map task is created per FileSplit.

When an individual map task starts it will open a new output writer per configured reduce task. It will then proceed to read its FileSplit using the RecordReader it gets from the specified InputFormat. InputFormat parses the input and generates key-value pairs. InputFormat must also handle records that may be split on the FileSplit boundary. For example TextInputFormat will read the last line of the FileSplit past the split boundary and, when reading other than the first FileSplit, TextInputFormat ignores the content up to the first newline.

Reference:How Map and Reduce operations are actually carried out

http://wiki.apache.org/hadoop/HadoopMapReduce(Map, second paragraph)

QUESTION NO: 5