CCD-410$^{Q\&As}$

Cloudera Certified Developer for Apache Hadoop (CCDH)

Pass Cloudera CCD-410 Exam with 100% Guarantee

QUESTION 1

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) ; ?

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 2

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 3

You need to create a job that does frequency analysis on input data. You will do this by writing a Mapper that uses TextInputFormat and splits each value (a line of text from an input file) into individual characters. For each one of these characters, you will emit the character as a key and an InputWritable as the value. As this will produce proportionally more intermediate data than input data, which two resources should you expect to be bottlenecks?

A. Processor and network I/O

B. Disk I/O and network I/O

C. Processor and RAM

D. Processor and disk I/O

Correct Answer: B

QUESTION 4

Which process describes the lifecycle of a Mapper?

A. The JobTracker calls the TaskTracker\\'s configure () method, then its map () method and finally its close () method.

B. The TaskTracker spawns a new Mapper to process all records in a single input split.

C. The TaskTracker spawns a new Mapper to process each key-value pair.

D. The JobTracker spawns a new Mapper to process all records in a single file.

Correct Answer: B

For each map instance that runs, the TaskTracker creates a new instance of your mapper. Note:

*

The Mapper is responsible for processing Key/Value pairs obtained from the InputFormat. The mapper may perform a number of Extraction and Transformation functions on the Key/Value pair before ultimately outputting none, one or many Key/Value pairs of the same, or different Key/Value type.

*

With the new Hadoop API, mappers extend the org.apache.hadoop.mapreduce.Mapper class. This class defines an \\'Identity\\' map function by default - every input Key/Value pair obtained from the InputFormat is written out.

Examining the run() method, we can see the lifecycle of the mapper:

```
/**

*

 Expert users can override this method for more complete control over the

*

 execution of the Mapper.

*

 @param context

*

 @throws IOException

*/

public void run(Context context) throws IOException, InterruptedException {

setup(context);

while (context.nextKeyValue()) {

map(context.getCurrentKey(), context.getCurrentValue(), context);

}

cleanup(context);

}
```

setup(Context) - Perform any setup for the mapper. The default implementation is a no-op method.

map(Key, Value, Context) - Perform a map operation in the given Key / Value pair. The default

implementation calls Context.write(Key, Value)

cleanup(Context) - Perform any cleanup for the mapper. The default implementation is a no-op method.

Reference: Hadoop/MapReduce/Mapper

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?