Exam Number/Code: CCD-333

Exam Name: Cloudera Certified Developer for Apache Hadoop

Version: Demo

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

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

Reference:http://stackoverflow.com/questions/9721754/how-to-parse-customwritable-from-text-inhadoop(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: B

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

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

QUESTION NO: 5
Which of the following statements most accurately describes the relationship between MapReduce and Pig?

A. Pig provides additional capabilities that allow certain types of data manipulation not possible with MapReduce.
B. Pig provides no additional capabilities to MapReduce. Pig programs are executed as MapReduce jobs via the Pig interpreter.
C. Pig programs rely on MapReduce but are extensible, allowing developers to do special-purpose processing not provided by MapReduce.
D. Pig provides the additional capability of allowing you to control the flow of multiple MapReduce jobs.

Answer: D

Reference:http://ofps.oreilly.com/titles/9781449302641/advanced_pig_latin.html(topic: controlling execution)

QUESTION NO: 6
You need to import a portion of a relational database every day as files to HDFS, and generate Java classes to Interact with your imported data. Which of the following tools should you use to accomplish this?

A. Pig
B. Hue
C. Hive
D. Flume
E. Sqoop
F. Oozie
G. fuse-dfs

Answer: C,E

Reference:http://log.medcl.net/item/2011/08/hadoop-and-mapreduce-big-data-analyticsgartner/(Data Movement between hadoop and relational databases, second paragraph)

QUESTION NO: 7
You have an employee who is a Date Analyst and is very comfortable with SQL. He would like to run ad-hoc analysis on data in your HDFS duster. Which of the following is a data warehousing software built on top of Apache Hadoop that defines a simple SQL-like query language well-suited for this kind of user?

A. Pig
B. Hue
C. Hive
D. Sqoop
E. Oozie
F. Flume
G. Hadoop Streaming

Answer: C

Reference:https://cwiki.apache.org/Hive/(Apache Hive, first sentence and second paragraph)

QUESTION NO: 8
Workflows expressed in Oozie can contain:

A. Iterative repetition of MapReduce jobs until a desired answer or state is reached.

B. Sequences of MapReduce and Pig jobs. These are limited to linear sequences of actions with exception handlers but no forks.
C. Sequences of MapReduce jobs only; no Pig or Hive tasks or jobs. These MapReduce sequences can be combined with forks and path joins.
D. Sequences of MapReduce and Pig. These sequences can be combined with other actions including forks, decision points, and path joins.

Answer: D

Reference:http://incubator.apache.org/oozie/docs/3.1.3/docs/WorkflowFunctionalSpec.html(workflow definition, first sentence)

QUESTION NO: 9
You need a distributed, scalable, data Store that allows you random, realtime read/write access to hundreds of terabytes of data. Which of the following would you use?

A. Hue
B. Pig
C. Hive
D. Oozie
E. HBase
F. Flume
G. Sqoop

Answer: E

Reference:http://hbase.apache.org/(when would I use HBase? First sentence)

QUESTION NO: 10
Which of the following utilities allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer?

A. Oozie
B. Sqoop
C. Flume
D. Hadoop Streaming

Answer: D

Reference:http://hadoop.apache.org/common/docs/r0.20.1/streaming.html(Hadoop Streaming, second sentence)

QUESTION NO: 11
What is the preferred way to pass a small number of configuration parameters to a mapper or reducer?

A. As key-value pairs in the jobconf object.
B. As a custom input key-value pair passed to each mapper or reducer.
C. Using a plain text file via the Distributedcache, which each mapper or reducer reads.
D. Through a static variable in the MapReduce driver class (i.e., the class that submits the MapReduce job).

Answer: B

QUESTION NO: 12
Given a Mapper, Reducer, and Driver class packaged into a jar, which is the correct way of submitting the job to the cluster?

A. jar MyJar.jar
B. jar MyJar.jar MyDriverClass inputdir outputdir
C. hadoop jar MyJar.jar MyDriverClass inputdir outputdir
D. hadoop jar class MyJar.jar MyDriverClass inputdir outputdir

Answer: C

Reference:http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0CGMQFjAC&url=http%3A%2F%2Fwww.cis.upenn.edu%2F~cis455%2Fslides%2F13-MapReduce-II.pptx&ei=ZwYpUJiyBoiM4gT4rYCQAg&usg=AFQjCNEFCHRaiW_a7QE9mpdwIGJ5OVmvA(slide 31, see step 2, run hadoop)

QUESTION NO: 13
What is the difference between a failed task attempt and a killed task attempt?

A. A failed task attempt is a task attempt that threw an unhandled exception. A killed task attempt is one that was terminated by the JobTracker.
B. A failed task attempt is a task attempt that did not generate any key value pairs. A killed task attempt is a task attempt that threw an exception, and thus killed by the execution framework.
C. A failed task attempt is a task attempt that completed, but with an unexpected status value. A killed task attempt is a duplicate copy of a task attempt that was started as part of speculative execution.
D. A failed task attempt is a task attempt that threw a RuntimeException (i.e., the task fails). A killed task attempt is a task attempt that threw any other type of exception (e.g.,

IOException); the execution framework catches these exceptions and reports them as killed.

Answer: D

QUESTION NO: 14
Custom programmer-defined counters in MapReduce are:

A. Lightweight devices for bookkeeping within MapReduce programs.
B. Lightweight devices for ensuring the correctness of a MapReduce program. Mappers Increment counters, and reducers decrement counters. If at the end of the program the counters read zero, then you are sure that the job completed correctly.
C. Lightweight devices for synchronization within MapReduce programs. You can use counters to coordinate execution between a mapper and a reducer.

Answer: B

Reference:http://hadooptutorial.wikispaces.com/Iterative+MapReduce+and+Counters(counters,second paragraph)

QUESTION NO: 15
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 file 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.

Answer: C