HADOOP-PR000007$^{Q\&As}$

Hortonworks Certified Apache Hadoop 2.0 Developer (Pig and Hive Developer)

Pass Hortonworks HADOOP-PR000007 Exam with 100% Guarantee

QUESTION 1

Given a directory of files with the following structure: line number, tab character, string: Example: 1abialkjfjkaoasdfjksdlkjhqweroij 2kadfjhuwqounahagtnbvaswslmnbfgy 3kjfteiomndscxeqalkzhtopedkfsikj 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

What data does a Reducer reduce method process?

A. All the data in a single input file.

B. All data produced by a single mapper.

C. All data for a given key, regardless of which mapper(s) produced it.

D. All data for a given value, regardless of which mapper(s) produced it.

Correct Answer: C

Explanation: Reducing lets you aggregate values together. A reducer function receives an iterator of input values from an input list. It then combines these values together, returning a single output value.

All values with the same key are presented to a single reduce task.

Reference: Yahoo! Hadoop Tutorial, Module 4: MapReduce

QUESTION 3

Your client application submits a MapReduce job to your Hadoop cluster. Identify the Hadoop daemon on which the Hadoop framework will look for an available slot schedule a MapReduce operation.

A. TaskTracker

B. NameNode

C. DataNode

D. JobTracker

E. Secondary NameNode

Correct Answer: D

Explanation: JobTracker is the daemon service for submitting and tracking MapReduce jobs in Hadoop. There is only One Job Tracker process run on any hadoop cluster. Job Tracker runs on its own JVM process. In a typical production cluster its run on a separate machine. Each slave node is configured with job tracker node location. The JobTracker is single point of failure for the Hadoop MapReduce service. If it goes down, all running jobs are halted. JobTracker in Hadoop performs following actions(from Hadoop Wiki:)

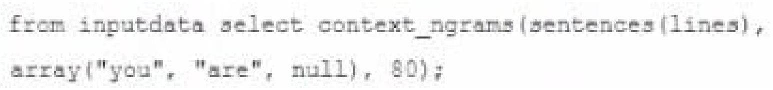
Client applications submit jobs to the Job tracker. The JobTracker talks to the NameNode to determine the location of the data The JobTracker locates TaskTracker nodes with available slots at or near the data The JobTracker submits the work to the chosen TaskTracker nodes. The TaskTracker nodes are monitored. If they do not submit heartbeat signals often enough, they are deemed to have failed and the work is scheduled on a different TaskTracker. A TaskTracker will notify the JobTracker when a task fails. The JobTracker decides what to do then: it may resubmit the job elsewhere, it may mark that specific record as something to avoid, and it may may even blacklist the TaskTracker as unreliable. When the work is completed, the JobTracker updates its status.

Client applications can poll the JobTracker for information.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers, What is a JobTracker in Hadoop? How many instances of JobTracker run on a Hadoop Cluster?

QUESTION 4

Assuming the following Hive query executes successfully:



```
from inputdata select context_ngrams(sentences(lines),
array("you", "are", null), 80);
```

Which one of the following statements describes the result set?

A. A bigram of the top 80 sentences that contain the substring "you are" in the lines column of the input data A1 table.

B. An 80-value ngram of sentences that contain the words "you" or "are" in the lines column of the inputdata table.

C. A trigram of the top 80 sentences that contain "you are" followed by a null space in the lines column of the inputdata table.

D. A frequency distribution of the top 80 words that follow the subsequence "you are" in the lines column of the inputdata table.

Correct Answer: D

QUESTION 5

Indentify the utility that 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

E. mapred

Correct Answer: D

Explanation: Hadoop streaming is a utility that comes with the Hadoop distribution. The utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

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

QUESTION 6

In the reducer, the MapReduce API provides you with an iterator over Writable values. What does calling the next () method return?

A. It returns a reference to a different Writable object time.

B. It returns a reference to a Writable object from an object pool.

C. It returns a reference to the same Writable object each time, but populated with different data.

D. It returns a reference to a Writable object. The API leaves unspecified whether this is a reused object or a new object.

E. It returns a reference to the same Writable object if the next value is the same as the previous value, or a new Writable object otherwise.

Correct Answer: C

Explanation: Calling Iterator.next() will always return the SAME EXACT instance of IntWritable, with the contents of that instance replaced with the next value.

Reference: manupulating iterator in mapreduce

QUESTION 7

Which one of the following statements is false about HCatalog?

A. Provides a shared schema mechanism

B. Designed to be used by other programs such as Pig, Hive and MapReduce

C. Stores HDFS data in a database for performing SQL-like ad-hoc queries

D. Exists as a subproject of Hive

Correct Answer: C

QUESTION 8

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 9

Which Hadoop component is responsible for managing the distributed file system metadata?

A. NameNode

B. Metanode

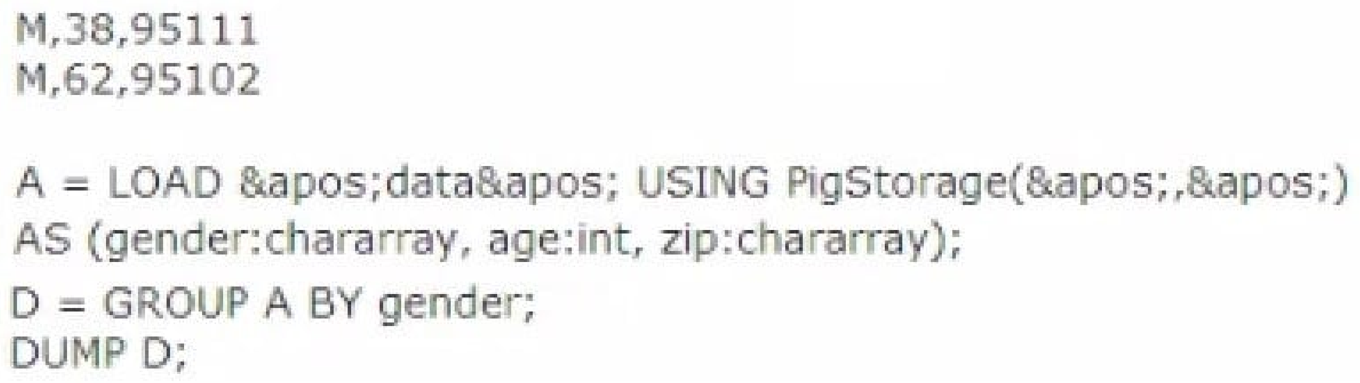
C. DataNode

D. NameSpaceManager

Correct Answer: A

QUESTION 10

Review the following andapos;dataandapos; file and Pig code.



```
M,38,95111
M,62,95102

A = LOAD &apos;data&apos; USING PigStorage(&apos;,&apos;)
AS (gender:chararray, age:int, zip:chararray);
D = GROUP A BY gender;
DUMP D;
```

Which one of the following statements is true?

A. The Output Of the DUMP D command IS (M,{(M,62.95102),(M,38,95111)})

B. The output of the dump d command is (M, {(38,95in),(62,95i02)})

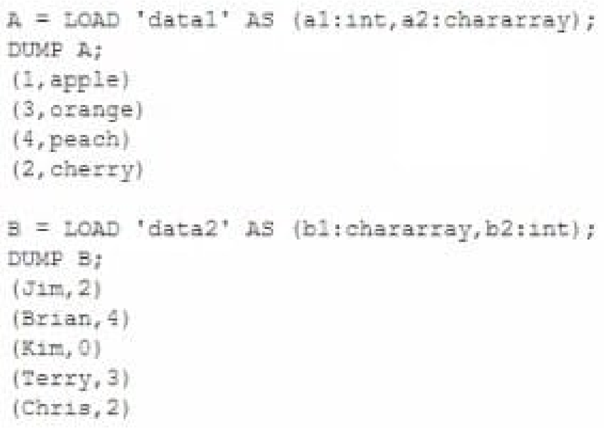
C. The code executes successfully but there is not output because the D relation is empty

D. The code does not execute successfully because D is not a valid relation

Correct Answer: A

QUESTION 11

Consider the following two relations, A and B.



```
A = LOAD 'data1' AS (a1:int,a2:chararray);
DUMP A;
(1,apple)
(3,orange)
(4,peach)
(2,cherry)

B = LOAD 'data2' AS (b1:chararray,b2:int);
DUMP B;
(Jim,2)
(Brian,4)
(Kim,0)
(Terry,3)
(Chris,2)
```

Which Pig statement combines A by its first field and B by its second field?

A. C = DOIN B BY a1, A by b2;

B. C = JOIN A by al, B by b2;

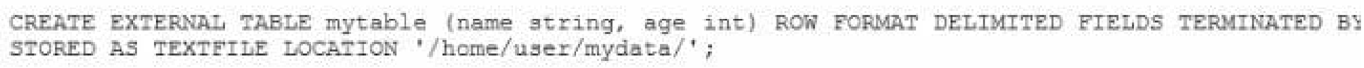
C. C = JOIN A a1, B b2;

D. C = JOIN A SO, B $1;

Correct Answer: B

QUESTION 12

Given the following Hive command:



```
CREATE EXTERNAL TABLE mytable (name string, age int) ROW FORMAT DELIMITED FIELDS TERMINATED BY
STORED AS TEXTFILE LOCATION '/home/user/mydata/';
```

Which one of the following statements is true?

A. The files in the mydata folder are copied to a subfolder of /apps/hlve/warehouse

B. The files in the mydata folder are moved to a subfolder of /apps/hive/wa re house

C. The files in the mydata folder are copied into Hive\\'s underlying relational database

D. The files in the mydata folder do not move from their current location In HDFS

Correct Answer: D

QUESTION 13

Analyze each scenario below and indentify which best describes the behavior of the default partitioner?

A. The default partitioner assigns key-values pairs to reduces based on an internal random number generator.

B. The default partitioner implements a round-robin strategy, shuffling the key-value pairs to each reducer in turn. This ensures an event partition of the key space.

C. The default partitioner computes the hash of the key. Hash values between specific ranges are associated with different buckets, and each bucket is assigned to a specific reducer.

D. The default partitioner computes the hash of the key and divides that valule modulo the number of reducers. The result determines the reducer assigned to process the key-value pair.

E. The default partitioner computes the hash of the value and takes the mod of that value with the number of reducers. The result determines the reducer assigned to process the key-value pair.

Correct Answer: D

Explanation: The default partitioner computes a hash value for the key and assigns the partition based on this result.

The default Partitioner implementation is called HashPartitioner. It uses the hashCode() method of the key objects modulo the number of partitions total to determine which partition to send a given (key, value) pair to.

In Hadoop, the default partitioner is HashPartitioner, which hashes a record\\'s key to determine which partition (and thus which reducer) the record belongs in.The number of partition is then equal to the number of reduce tasks for the job.

Reference: Getting Started With (Customized) Partitioning

QUESTION 14

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

Correct Answer: B

Reference: Hadoop binary files processing introduced by image duplicates finder

QUESTION 15

Which one of the following statements describes a Hive user-defined aggregate function?

A. Operates on multiple input rows and creates a single row as output

B. Operates on a single input row and produces a single row as output

C. Operates on a single input row and produces a table as output

D. Operates on multiple input rows and produces a table as output

Correct Answer: A