Vendor: Cloudera

Exam Code:DS-200

Exam Name: Data Science Essentials

Version: Demo

QUESTION 1

In what way can Hadoop be used to improve the performance of LIoyd\\'s algorithm for k-means clustering on large data sets?

A. Parallelizing the centroid computations to improve numerical stability

B. Distributing the updates of the cluster centroids

C. Reducing the number of iterations required for the centroids to converge

D. Mapping the input data into a non-Euclidean metric space

Correct Answer: B

QUESTION 2

You\\'ve built a model that has ten different variables with complicated independence relationships between them, and both continuous and discrete variables that have complicated, multi-parameter distributions. Computing the joint probability distribution is complex, but it turns out that computing the conditional probabilities for the variables is easy. What is the most computationally efficient for computing the expected value?

A. Method of moments

B. Markov Chain Monte Carlo

C. Gibbs sampling

D. Numerical quadrature

Correct Answer: B

QUESTION 3

You have just run a MapReduce job to filter user messages to only those of a selected geographical region. The output for this job in a directory named westUsers, located just below your home directory in HDFS. Which command gathers these records into a single file on your local file system?

A. Hadoop fs getmerge westUsers WestUsers.txt

B. Hadoop fs get westUsers WestUsers.txt

C. Hadoop fs cp westUsers/* westUsers.txt

D. Hadoop fs getmerge R westUsers westUsers.txt

Correct Answer: B

QUESTION 4

There are 20 patients with acute lymphoblastic leukemia (ALL) and 32 patients with acute myeloid leukemia (AML), both variants of a blood cancer.

The makeup of the groups as follows:

ALL GROUP			
	Male	Female	
Caucasian	14	1	15
Asian-American	5	0	5
	19	1	20

AML GROUP			
	Male	Female	
Caucasian	9	4	13
Asian-American	7	12	19
	16	16	32

Each individual has an expression value for each of 10000 different genes. The expression value for each gene is a continuous value between -1 and 1.

You want to use the data from the 52 patients in the scenario to improve the ability of doctors being able to distinguish between ALL and AML. What type of data science problem is this?

A. Classification

B. Regression

C. Clustering

D. Filtering

Correct Answer: D

QUESTION 5

You are about to sample a 100-dimensinal unit-cube. To adequately sample any single given dimension, you need only capture 10 points. How many points do you need to order to sample the complete 100dimensional unit cube adequately?

A. 10010

B. 1010

C. Log2(100)

D. 100

E. 1000

F. 1010

Correct Answer: E

QUESTION 6

What is default delimiter for Hive tables?

A. ^A (Control-A)

B. , (comma)

C. \t (tab)

D. : (colon)

Correct Answer: A

Reference: http://blog.spryinc.com/2013/10/four-useful-tricks-for-working-with-hive.html (change the delimiter when exporting hive table)

QUESTION 7

Given the following sample of numbers from a distribution:

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89 What are the five numbers that summarize this distribution (the five number summary of sample percentiles)?

A. 1, 3, 8, 34, 89

B. 1, 4, 13, 34, 89

C. 1, 1.5, 5, 24.5, 89

D. 1, 2.5, 8, 27.5, 89

Correct Answer: A

QUESTION 8

You need to analyze 60,000,000 images stored in JPEG format, each of which is approximately 25 KB. Because your Hadoop cluster isn\\'t optimized for storing and processing many small files you decide to do the following actions:

1.

Group the individual images into a set of larger files

2.

Use the set of larger files as input for a MapReduce job that processes them directly with Python using Hadoop streaming

Which data serialization system gives you the flexibility to do this?

A. CSV

B. XML

C. HTML

D. Avro

E. Sequence Files

F. JSON

Correct Answer: BF

QUESTION 9

You have acquired a new data source of millions of customer records, and you\\'ve this data into HDFS. Prior to analysis, you want to change all customer registration to the same date format, make all addresses uppercase, and remove all customer names (for anonymization). Which process will accomplish all three objectives?

A. Adapt the data cleansing module in Mahout to your data, and invoke the Mahout library when you run your analysis

B. Pull this data into an RDBMS using sqoop and scrub records using stored procedures

C. Write a script that receives records on stdin, corrects them, and then writes them to stdout. Then, invoke this script in a map-only Hadoop Streaming Job

D. Write a MapReduce job with a mapper to change words to uppercase and to reduce different forms of dates to a single form

Correct Answer: C

QUESTION 10

You have a large m x n data matrix M. You decide you want to perform dimension reduction/clustering on your data and

have decide to use the singular value decomposition (SVD; also called principal components analysis PCA)

Refer to the passage above.

What represents the SVD of the Matrix standard M given the following information:

U is m x m unitary V is n x n unitary S is m x n diagonal Q is n x n invertible D is n x n diagonal L is m x m lower triangular U is m x m upper triangular

A. M = U S V

B. M = U P

C. M = Q D Q-1

D. M = L U

Correct Answer: A

QUESTION 11

You are building a k-nearest neighbor classifier (k-NN) on a labeled set of points in a high- dimensional space. You determine that the classifier has a large error on the training data. What is the most likely problem?

A. High-dimensional spaces effectively make local neighborhoods global

B. k-NN compotation does not coverage in high dimensions

C. k was too small

D. The VC-dimension of a k-NN classifier is too high

Correct Answer: B

QUESTION 12

You have user profile records in an OLTP database that you want to join with web server logs which you have already ingested into HDFS. What is the best way to acquire the user profile for use in HDFS?

A. Ingest with Hadoop streaming

B. Ingest with Apache Flume

C. Ingest using Hive\\'s LOAD DATA command

D. Ingest using Sqoop

E. Ingest using Pig\\'s LOAD command

Correct Answer: BD

Reference: https://thinkbiganalytics.com/leading_big_data_technologies/ingestion-and- streaming-withstorm-kafka-flume/