DS-200 Q&As

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.passapply.com/ds-200.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Cloudera Official Exam Center

QUESTION 1

You want to understand more about how users browse your public website. For example, you war know which pages they visit prior to placing an order. You have a server farm of 200 web server hosting your website. Which is the most efficient process to gather these web servers access logs into your Hadoop cluster for analysis?

A. Sample the web server logs web servers and copy them into HDFS using curl

B. Channel these click streams into Hadoop using Hadoop Streaming

C. Write a MapReduce job with the web servers for mappers and the Hadoop cluster nodes for reducers

D. Import all user clicks from your OLTP databases Into Hadoop using Sqoop

E. Ingest the server web logs into HDFS using Flume

Correct Answer: C

QUESTION 2

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/

QUESTION 3

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 4

Given the following sample of numbers from a distribution: 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89

What are two benefits of using the five-number summary of sample percentiles to summarize a data set?

A. You can calculate unbiased estimators for the parameters of the distribution

B. It\\'s robust to outliers

C. It\\'s well-defined for any probability distribution

D. You can calculate it quickly using a relational database like MySQL, even when we have a large sample

Correct Answer: D

QUESTION 5

Which two machine learning algorithm should you consider as likely to benefit from discretizing continuous features?

A. Support vector machine

B. Naïve Bayes

C. Decision trees

D. Logistic regression

E. Singular value decomposition

Correct Answer: AB

Reference: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656082/

QUESTION 6

You are building a system to perform outlier detection for a large online retailer. You need to build a system to detect if the total dollar value of sales are outside the norm for each U.S. state, as determined from the physical location of the buyer for each purchase. The retailer\\'s data sources are scattered across multiple systems and databases and are unorganized with little coordination or shared data or keys between the various data sources.

Below are the sources of data available to you. Determine which three will give you the smallest set of data sources but still allow you to implement the outlier detector by state.

A. Database of employees that Includes only the employee ID, start date, and department

B. Database of users that contains only their user ID, name, and a list of every Item the user has viewed

C. Transaction log that contains only basket ID, basket amount, time of sale completion, and a session ID

D. Database of user sessions that includes only session ID, corresponding user ID, and the corresponding IP address

E. External database mapping IP addresses to geographic locations

F. Database of items that includes only the item name, item ID, and warehouse location

G. Database of shipments that includes only the basket ID, shipment address, shipment date, and shipment method

Correct Answer: ADF

QUESTION 7

You have a large file of N records (one per line), and want to randomly sample 10% them. You have two

functions that are perfect random number generators (through they are a bit slow):

Random_uniform () generates a uniformly distributed number in the interval [0, 1] random_permotation (M)

generates a random permutation of the number O through M -1.

Below are three different functions that implement the sampling.

Method A

For line in file: If random_uniform ()

Method B

i = 0

for line in file:

if i % 10 = = 0;

print line

i += 1

Method C

idxs = random_permotation (N) [: (N/10)]

i = 0

for line in file:

if i in idxs:

print line

i +=1

Which method might introduce unexpected correlations?

A. Method A

B. Method B

C. Method C

Correct Answer: C

QUESTION 8

You have a large file of N records (one per line), and want to randomly sample 10% them. You have two functions that are perfect random number generators (through they are a bit slow):

Random_uniform () generates a uniformly distributed number in the interval [0, 1] random_permotation (M) generates a random permutation of the number O through M -1.

Below are three different functions that implement the sampling.

Method A

For line in file: If random_uniform ()

Method B

i = 0

for line in file:

if i % 10 = = 0;

print line

i += 1

Method C

idxs = random_permotation (N) [: (N/10)]

i = 0

for line in file:

if i in idxs:

print line

i +=1

Which method is least likely to give you exactly 10% of your data?

A. Method A

B. Method B

C. Method C

Correct Answer: B

QUESTION 9

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.

With which type of plot can you encode the most amount of the data visually?

You choose to perform agglomerative hierarchical clustering on the 10,000 features. How much RAM do you need to hold the distance Matrix, assuming each distance value is 64-bit double?

A. ~ 800 MB

B. ~ 400 MB

C. ~ 160 KB

D. ~ 4 MB

Correct Answer: B

QUESTION 10

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