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:

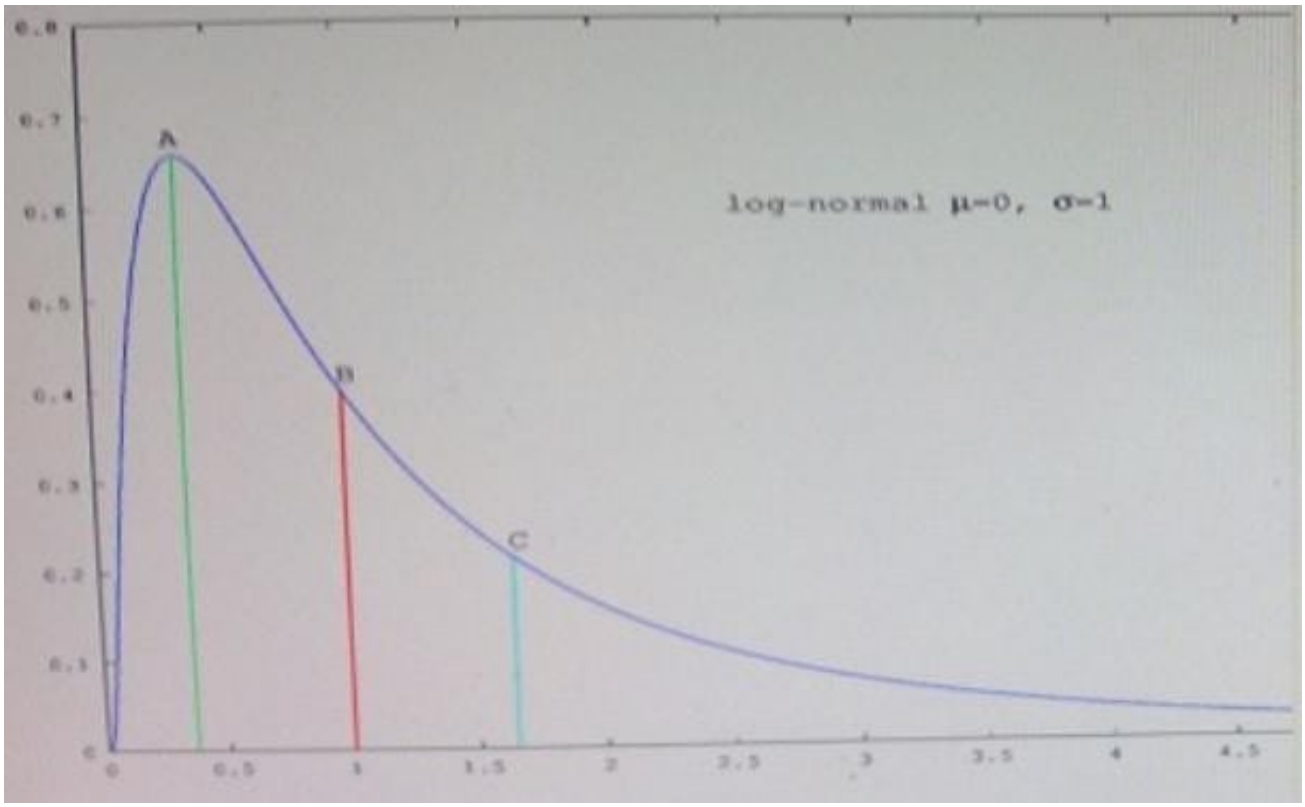
https://www.passapply.com/ds-200.html

100% Passing Guarantee
100% Money Back Assurance

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

QUESTION 1

Refer to the exhibit.



Which point in the figure is the mean?

A. A

B. B

C. C

Correct Answer: B

QUESTION 2

Which best describes the primary function of Flume?

A. Flume is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with an infrastructure consisting of sources and sinks for importing and evaluating large data sets

B. Flume acts as a Hadoop filesystem for log files

C. Flume Imports data from SQL/relational database into your Hadoop cluster

D. Flume provides a query languages for Hadoop similar to SQL

E. Flume is a distributed server for collecting and moving large amount of data into HDFS as it\\'s produced from streaming data flows

Correct Answer: D

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\\'ve built your model for discriminating between AML and ALL patients and you find that it works quite well on your current data. One month later, a collaboration tells you she has fresh data from 100 new AML/ ALL patients. You run the samples through your model, and turns out your model has very poor predictive accuracy on the new samples; specifically, your model predicts that all males have ALL. What is the most reliable way to fix this problem?

A. Change the distance metric

B. Reduce the number of dimensions

C. Use a Gibbs sampler on a Bayesian network

D. Perform matched sampling across other provided variables

Correct Answer: D

QUESTION 5

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