



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.geekcert.com/ds-200.html>

100% Passing Guarantee
100% Money Back Assurance

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

-  **Instant Download** After Purchase
-  **100% Money Back** Guarantee
-  **365 Days** Free Update
-  **800,000+** Satisfied Customers





QUESTION 1

When optimizing a function using stochastic gradient descent, how frequently should you update your estimate of the gradient?

- A. Once after every pass through the data set
- B. Once per observation
- C. For each observation with a probability that you choose ahead of time
- D. After a random number of observations
- E. Once every N observations, where you decide N ahead of time

Correct Answer: AC

QUESTION 2

You are working with a logistic regression model to predict the probability that a user will click on an ad. Your model has hundreds of features, and you're not sure if all of those features are helping your prediction. Which regularization technique should you use to prune features that aren't contributing to the model?

- A. Convex
- B. Uniform
- C. L2
- D. L1

Correct Answer: A

QUESTION 3

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 4

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_permutation (M) generates a random permutation of the number 0 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_permutation (N) [: (N/10)]
```

```
i = 0
```

```
for line in file:
```

```
if i in idxs:
```

```
print line
```

```
i +=1
```



Which method will have the best runtime performance?

- A. Method A
- B. Method B
- C. Method C

Correct Answer: A

QUESTION 5

Which three metrics are useful in measuring the accuracy and quality of a recommender system?

- A. Mutual Information
- B. RMSF
- C. Tanimoto coefficient
- D. Pearson correlation
- E. Precision
- F. Recall

Correct Answer: CDE

Reference: <https://lirias.kuleuven.be/bitstream/123456789/289803/3/datasets-cameraready.pdf>

[DS-200 VCE Dumps](#)

[DS-200 Study Guide](#)

[DS-200 Braindumps](#)