Trạng thái	Đã xong
Bắt đầu vào lúc	Thứ Ba, 9 tháng 12 2025, 8:13 AM
Kết thúc lúc	Thứ Ba, 9 tháng 12 2025, 8:34 AM
Thời gian thực hiện	21 phút 5 giây
Điểm	**25,00** trên 30,00 (**83,33**%)

Câu hỏi **1**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời chính xác về trực quan hóa (visualization) dữ liệu lớn:

a. Thách thức là sự lộn xộn về thị giác (visual clutter), các vấn đề về hiệu suất, nhận thức hạn chế. ✓
b. Cả A, B và C đều đúng.
c. Không có kỹ thuật trực quan nào hữu ích cho việc truyền dữ liệu.
d. Khai phá dữ liệu (Data mining), Mã hóa (Encoding) & Bố cục (Layout) và Kết xuất (Rendering) là vô dụng.

Câu hỏi **2**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời đúng về Spark:

a. Spark bao gồm các chức năng được sử dụng trong Data streaming, Học máy, đồ thị và SQL. ✓
b. Spark không thể chạy một mình mà phải chạy trên Hadoop.
c. Spark là một công cụ để lưu trữ dữ liệu lớn vào đĩa cứng như Hadoop nhưng hiệu suất của nó tốt hơn.
d. Cả A, B và C đều đúng.

Câu hỏi **3**

Sai

Đạt điểm 0,00 trên 1,00

Select one:

a. ✗
b.
c.
d.

Câu hỏi **4**

Đúng

Đạt điểm 1,00 trên 1,00

Chuỗi bước điển hình để phát hiện gần trùng lặp văn bản bằng LSH là gì?

Select one:

a. PCA → k-d tree → Tính Jaccard chính xác

b. Tokenization → TF–IDF → K-means

c. Loại bỏ stopword → Word2Vec → HNSW

d. Shingling → MinHash → LSH (banding) ✓

Câu hỏi **5**

Đúng

Đạt điểm 1,00 trên 1,00

Thiết kế pipeline đơn giản cho ứng dụng giám sát giao dịch ngân hàng theo thời gian thực:

a. Dữ liệu -> Kafka -> Spark Streaming -> Dashboard ✓

b. Dữ liệu -> MySQL -> Spark

c. Dữ liệu -> Hadoop -> Spark Batch

d. Dữ liệu -> Excel -> PowerBI

Câu hỏi **6**

Sai

Đạt điểm 0,00 trên 1,00

Trong LSH cho khoảng cách Euclid với phép chiếu lên trục ngẫu nhiên và lượng tử hóa theo bề rộng ‘bucket’ a, trực giác đúng là:

Select one:

a. ✗

b.

c.

d.

Câu hỏi **7**

Đúng

Đạt điểm 1,00 trên 1,00

Mục tiêu chính của hàm mục tiêu trong thuật toán K-means là gì?

a. Tối đa hóa độ tương đồng giữa các điểm trong các cụm khác nhau
b. Tối đa hóa khoảng cách giữa các tâm cụm
c. Tối thiểu hóa số lượng cụm được tạo ra
d. Tối thiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm của chúng ✓

Câu hỏi **8**

Sai

Đạt điểm 0,00 trên 1,00

Chọn câu trả lời đúng về K-means:

a. Không có gì được đề cập.
b. K-means là một thuật toán có chi phí tính toán nhỏ cho kích thước bài toán lớn.
c. Không thể sử dụng phép đo khoảng cách Jaccard trong K-means. ✗
d. K-means có thể là một giải pháp trong việc phân cụm mạng xã hội.

Câu hỏi **9**

Đúng

Đạt điểm 1,00 trên 1,00

Trong Hadoop, HDFS có vai trò gì?

a. Quản lý tác vụ xử lý dữ liệu
b. Thực hiện tính toán MapReduce
c. Phân phối yêu cầu đến các consumer
d. Lưu trữ dữ liệu phân tán trên nhiều nút (nodes) ✓

Câu hỏi **10**

Đúng

Đạt điểm 1,00 trên 1,00

4V của dữ liệu lớn là:

a. Khối lượng, Vận tốc, Sự đa dạng và Minh bạch.
b. Khối lượng, Vận tốc, Sự đa dạng và Tính xác thực. ✓
c. Khối lượng, vận tốc, đa dạng và nhất quán.
d. Không có gì được đề cập.

Câu hỏi **11**

Đúng

Đạt điểm 1,00 trên 1,00

Trong hệ thống sử dụng đồng hồ logic Lamport, nếu sự kiện A có timestamp C(A) = 5 và sự kiện B có timestamp C(B) = 8, chúng ta có thể kết luận chắc chắn điều gì?

Select one:

a. Sự kiện A chắc chắn đã xảy ra trước sự kiện B.

b. Sự kiện A và B chắc chắn xảy ra đồng thời.

c. Hệ thống đang bị lỗi vì timestamp quá chênh lệch.

d. Sự kiện B không thể xảy ra trước sự kiện A. ✓

Câu hỏi **12**

Đúng

Đạt điểm 1,00 trên 1,00

Phát hiện cộng đồng (community detection) trong đồ thị mạng xã hội:

a. PageRank được cá nhân hóa được sử dụng để tính toán Conductance cho một cộng đồng.

b. Tất cả những gì đã đề cập. ✓

c. K-means là một giải pháp tốt.

d. Cộng đồng được phát hiện có thể có độ dẫn điện tối thiểu cục bộ tương ứng với các cụm tốt.

Câu hỏi **13**

Đúng

Đạt điểm 1,00 trên 1,00

Thuật toán PageRank được phát triển ban đầu bởi ai?

a. Mark Zuckerberg và Eduardo Saverin

b. Steve Jobs và Steve Wozniak

c. Larry Page và Sergey Brin ✓

d. Bill Gates và Paul Allen

Câu hỏi **14**

Đúng

Đạt điểm 1,00 trên 1,00

Phát biểu nào sau đây là ĐÚNG về việc khởi tạo các tâm cụm (centroids) trong K-means?

a. Vị trí khởi tạo không ảnh hưởng đến tốc độ hội tụ của thuật toán
b. Các tâm cụm bắt buộc phải được chọn từ các điểm dữ liệu có sẵn
c. Việc chọn vị trí khởi tạo ngẫu nhiên có thể dẫn đến các kết quả hội tụ khác nhau ✓
d. Thuật toán luôn hội tụ về cùng một kết quả bất kể khởi tạo thế nào

Câu hỏi **15**

Đúng

Đạt điểm 1,00 trên 1,00

Tại sao các liên kết nội bộ (internal links) cũng quan trọng trong PageRank?

a. Vì chúng làm giảm tỷ lệ thoát trang
b. Vì Google cấm sử dụng liên kết ra ngoài
c. Vì chúng giúp phân phối dòng chảy PageRank (link juice) đến các trang quan trọng khác trong cùng website ✓
d. Vì liên kết nội bộ luôn có giá trị cao hơn liên kết từ bên ngoài (backlinks)

Câu hỏi **16**

Sai

Đạt điểm 0,00 trên 1,00

Những đặc điểm nào KHÔNG phải trong data streaming:

a. Bộ nhớ hạn chế.
b. Khái niệm trên đường chuyền và trôi dạt.
c. Sức mạnh tính toán hạn chế. ✗
d. Truy vấn liên tục nhưng không truy vấn chuyên biệt (ad-hoc queries).

Câu hỏi **17**

Đúng

Đạt điểm 1,00 trên 1,00

Trong việc tìm kiếm các mục tài liệu tương tự:

a. Một giải pháp kết hợp Min-hashing và LSH giúp giảm chi phí tính toán.
b. Tất cả những gì đã đề cập. ✓
c. Băm tối thiểu (Min-hashing) được sử dụng như một giải pháp giảm kích thước.
d. Băm nhạy cảm cục bộ (LSH) có thể đề xuất các mục ít nhất là ngưỡng tương tự do người dùng xác định.

Câu hỏi **18**

Đúng

Đạt điểm 1,00 trên 1,00

Giả sử dữ liệu từ Kafka đến Spark không theo thứ tự thời gian, bạn nên dùng cơ chế nào của Spark để đảm bảo tính đúng đắn khi tính toán theo cửa sổ thời gian (window)?

a. Replication
b. Trigger Once
c. Partition Rebalance
d. Watermarking ✓

Câu hỏi **19**

Đúng

Đạt điểm 1,00 trên 1,00

Nếu bạn cần chạy một tác vụ tính tổng giá trị giao dịch từ hàng tỷ bản ghi, bạn nên sử dụng thành phần nào của Hadoop?

a. MapReduce ✓
b. YARN
c. Hive
d. HDFS

Câu hỏi **20**

Đúng

Đạt điểm 1,00 trên 1,00

Thành phần chính đầu tiên (First Principal Component) trong PCA có đặc điểm gì?

a. Luôn trùng với trục hoành (trục x) của hệ tọa độ gốc
b. Là hướng mà dữ liệu có phương sai (variance) lớn nhất ✓
c. Là hướng vuông góc với mọi thành phần chính khác
d. Là hướng có phương sai nhỏ nhất để giảm nhiễu

Câu hỏi **21**

Sai

Đạt điểm 0,00 trên 1,00

Chọn câu trả lời đúng về MapReduce:

a. Số lượng công nhân được sử dụng trong bước bản đồ thường lớn hơn trong bước giảm. ✗

b. Tất cả những gì đã đề cập.

c. Lúc đầu, các tác vụ tính toán được đặt trên các nút tính toán và sau đó dữ liệu được chuyển từ hệ thống tệp phân tán sang chúng.

d. Các lập trình viên có thể ghi đè lên hai hàm Map() và Reduce(), trong khi Combine() và Partition() được tạo tự động bởi MapReduce (hoặc có thể tùy biến).

Câu hỏi **22**

Đúng

Đạt điểm 1,00 trên 1,00

Vai trò của hệ số tắt dần (damping factor) trong công thức PageRank là gì?

a. Để loại bỏ các trang web spam khỏi kết quả tìm kiếm

b. Để mô phỏng xác suất người dùng tiếp tục nhấp vào liên kết thay vì tắt trình duyệt hoặc nhập URL mới ✓

c. Để tăng tốc độ tính toán của máy chủ

d. Để đảm bảo thuật toán dừng lại sau 10 bước

Câu hỏi **23**

Đúng

Đạt điểm 1,00 trên 1,00

Phân tích mối quan hệ giữa NameNode và DataNode trong HDFS:

a. NameNode lưu metadata, còn DataNode lưu dữ liệu thực tế ✓

b. Cả hai đều lưu trữ dữ liệu như nhau

c. DataNode điều phối nhiệm vụ cho NameNode

d. NameNode là node sao lưu dự phòng

Câu hỏi **24**

Đúng

Đạt điểm 1,00 trên 1,00

Trong phân rã SVD, ma trận $\sum(\sigma)$ có đặc điểm gì?

a. Là ma trận tam giác trên chứa các giá trị riêng (eigenvalues)
b. Là ma trận đường chéo chứa các giá trị kỳ dị (singular values) không âm ✓
c. Là ma trận vuông trực giao
d. Là ma trận chứa các vectơ riêng (eigenvectors) của A

Câu hỏi **25**

Đúng

Đạt điểm 1,00 trên 1,00

So sánh SVD (Singular Value Decomposition) và CUR trong giảm chiều dữ liệu:

a. CUR luôn cho kết quả tốt hơn SVD.
b. SVD tốt hơn CUR vì hiệu quả và thời gian chạy ngắn.
c. Cả A, B và C đều sai.
d. CUR có thời gian chạy ngắn và sử dụng ít bộ nhớ hơn SVD (đối với ma trận thưa lớn). ✓

Câu hỏi **26**

Đúng

Đạt điểm 1,00 trên 1,00

Kết hợp băm tối thiểu và băm nhạy cảm cục bộ (LSH) có thể được sử dụng trong:

a. Tất cả những gì đã đề cập. ✓
b. Đo độ giống nhau của khuôn mặt.
c. Hệ thống đề xuất.
d. Phát hiện đối tượng.

Câu hỏi **27**

Đúng

Đạt điểm 1,00 trên 1,00

Hadoop là gì?

a. Công cụ trực quan hóa dữ liệu
b. Hệ quản trị cơ sở dữ liệu quan hệ (RDBMS)
c. Framework xử lý dữ liệu lớn (Big Data) theo mô hình lập trình phân tán ✓
d. Hệ thống quản lý bộ nhớ

Câu hỏi **28**

Đúng

Đạt điểm 1,00 trên 1,00

Ưu điểm chính của **Vector Clock** so với Lamport Timestamp là gì?

Select one:

a. Loại bỏ hoàn toàn nhu cầu đồng bộ hóa đồng hồ vật lý.

b. Sử dụng ít không gian lưu trữ hơn trong mỗi thông điệp.

c. Cho phép xác định chính xác mối quan hệ nhân quả giữa hai sự kiện. ✓

d. Đơn giản hơn trong việc triển khai và tính toán.

Câu hỏi **29**

Đúng

Đạt điểm 1,00 trên 1,00

Ý tưởng cốt lõi đằng sau thuật toán PageRank là gì?

a. Đếm số lượng từ khóa xuất hiện trên trang

b. Xem các liên kết (hyperlinks) như những phiếu bầu độ tin cậy ✓

c. Dựa vào tốc độ tải trang của website

d. Dựa vào lưu lượng truy cập (traffic) hàng ngày của website

Câu hỏi **30**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời đúng về Machine Learning (ML) trong dữ liệu lớn:

a. Cả A, B và C đều sai. ✓

b. ML quá phức tạp để áp dụng dữ liệu lớn.

c. Hiện tại không có phần cứng nào hỗ trợ ML được sử dụng trong các bài toán dữ liệu lớn.

d. ML không hiệu quả vì quá nhiều dữ liệu đầu vào.