Câu 1: Đâu là kỹ thuật có thể được dùng để thích nghi các giải thuật học máy cho dữ liệu lớn?
A. Sub-sampling, principal component analysis, feature extraction và feature selection
B. Song song hóa trên Mapreduce hay Spark
C. Các kiến trúc mới xử lý luồng liên tục như mini-batch, complex event processing
D. Tất cả các ý (1), (2), (3)
E. Các ý (2) và (3)

Câu 2: Các mục tiêu chính của Apache Hadoop?
A. Lưu trữ dữ liệu khả mở
B. Xử lý dữ liệu lớn mạnh mẽ
C. Trực quan hoá dữ liệu hiệu quả
D. Lưu trữ dữ liệu khả mở và Xử lý dữ liệu lớn mạnh mẽ
E. Lưu trữ dữ liệu khả mở, xử lý dữ liệu lớn mạnh mẽ và trực quan hoá dữ liệu hiệu quả

Câu 3: Phát biểu nào sau đây không đúng về Apache Hadoop?
A. Xử lý dữ liệu phân tán với mô hình lập trình đơn giản, thân thiện hơn như MapReduce.
B. Hadoop thiết kế để mở rộng thông qua kỹ thuật scale-out, tăng số lượng máy chủ
C. Thiết kế để vận hành trên phần cứng phổ thông, có khả năng chống chịu lỗi phần cứng
D. Thiết kế để vận hành trên siêu máy tính, cấu hình mạnh, độ tin cậy cao

Câu 4: Thành phần nào không thuộc thành phần lõi của Hadoop?
A. Hệ thống tệp tin phân tán HDFS
B. Mapreduce framework
C. YARN: yet another resource negotiator
D. Apache Zookeeper
E. Apache Hbase

Câu 5: Hadoop giải quyết bài toán khả mở bằng cách nào? Chọn đáp án sai.
A. Thiết kế hướng phân tán ngay từ đầu, mặc định triển khai trên cụm máy chủ
B. Các node tham gia vào cụm Hadoop được gán vai trò hoặc là node tính hoán hoặc là node lưu trữ dữ liệu
C. Các node tham gia vào cụm đóng cả 2 vai trò tính toán và lưu trữ
D. Các node thêm vào cụm cần có cấu hình, độ tin cậy cao

Câu 6: Hadoop giải quyết bài toán chịu lỗi thông qua kỹ thuật gì. Chọn đáp án sai.
A. Hadoop chịu lỗi thông qua kỹ thuật dư thừa
B. Các tệp tin được phân mảnh, các mảnh được nhân bản ra các node khác trên cụm
C. Các tệp tin được phân mảnh, các mảnh được lưu trữ tin cậy trên ổ cứng theo cơ chế RAID
D. Các công việc cần tính toán được phân mảnh thành các tác vụ độc lập.

Câu 7: Công cụ nào có thể sử dụng để hỗ trợ import, export dữ liệu vào ra hệ sinh thái Hadoop?
A. Oozie (lên lịch và quản lý) phân lung, lp lch lung công vic
B. Flume (thu thập dữ liệu)
C. Sqoop (trung gian tương tác với SQL) H tr import, export
D. Hive (truy vấn) truy vn HIVEQL like SQL

Câu 8: Vai trò của YARN?
A. Quản lý và phân phối tài nguyên trong cụm Hadoop

B. Cung cấp giao diện người dùng mức cao, biến đổi truy vấn thành các job Mapreduce
C. Cung cấp các chức năng phối hợp phân tán độ tin cậy cao như quản lý thành viên, bầu cử, giám sát trạng thái hệ thống

Câu 9: Giữa Pig và Hive, công cụ nào có giao diện truy vấn gần với ANSI SQL hơn?
A. Pig
B. Hive
C. Pig và Hive đều không có giao diện truy vấn gần với SQL.

Câu 10: Siêu dữ liệu (metadata) trong hệ thống quản lý tệp tin là gì?
A. Là các tệp tin kích thước lớn hơn nhiều kích thước tệp tin phổ biến (từ vài GB tới TB).
B. Là dữ liệu mô tả về tệp tin bao gồm thời gian khởi tạo, thông tin phân quyền người dùng
C. Ánh xạ từ tệp tin tới vị trí lưu trữ vật lý của tệp tin.

Câu 11: Ưu điểm của hệ thống tệp tin phân tán là gì?
A. Đơn giản hoá việc chia sẻ dữ liệu.
B. Tập trung hoá việc quản trị.
C. Cho phép người dùng có cái nhìn hợp nhất (như nhau) về toàn bộ dữ liệu trong hệ thống

Câu 12: Ưu điểm của kiến trúc NAS (Network attached Storage)?
A. Đơn giản hoá việc chia sẻ dữ liệu.
B. Tính khả mở cao.
C. Máy khách có thể kết nối tới NAS bằng đường truyền Ethernet thông thường (Chuẩn kết nối TCP/IP).

Câu 13: Ưu điểm của kiến trúc SAN (Storage area network)?
A. Quản trị dễ dàng hơn so với NAS.
B. Hiệu năng, băng thông tốt hơn với NAS.
C. Máy khách có thể kết nối tới SAN bằng đường truyền Ethernet thông thường (Chuẩn kết nối TCP/IP)

Câu 14: Thế nào là UNIX semantic?
A. Cập nhật tới tệp tin có thể được nhìn thấy ngay lập tức bởi các tiến trình khác mà mở tệp tin đó cùng thời điểm với tiến trình ghi.
B. Tệp tin là chỉ đọc, không cho phép cập nhật và ghi đè. Mọi tiến trình đều có thể đọc tệp tin đồng thời
C. Cập nhật tới tệp tin chỉ có thể thấy được bởi các tiến trình khác sau khi tiến trình ghi thực hiện thao tác đóng tệp.

Câu 15: Thế nào là Session semantic?
A. Cập nhật tới tệp tin có thể được nhìn thấy ngay lập tức bởi các tiến trình khác mà mở tệp tin đó cùng thời điểm với tiến trình ghi.
B. Tệp tin là chỉ đọc, không cho phép cập nhật và ghi đè. Mọi tiến trình đều có thể đọc tệp tin đồng thời.
C. Cập nhật tới tệp tin chỉ có thể thấy được bởi các tiến trình khác sau khi tiến trình ghi thực hiện thao tác đóng tệp.

Câu 16: Các đặc trưng của HDFS. Chọn đáp án sai.
A. Tối ưu cho các tệp tin có kích thước lớn
B. Hỗ trợ thao tác đọc ghi tương tranh tại chunk (phân mảnh) trên tệp tin
C. Hỗ trợ nén dữ liệu để tiết kiệm chi phí

D. Hỗ trợ cơ chế phân quyền và kiểm soát người dùng của UNIX

Câu 17: Mô tả cách thức một client đọc dữ liệu trên HDFS.
A. Client truy vấn Namenode để biết được vị trí các chunks. Namenode trả về vị trí các chunks. Client kết nối song song tới các datanode để đọc các chunk
B. Client thông báo tới namenode để bắt đầu quá trình đọc sau đó client truy vấn các datanode để trực tiếp đọc các chunks
C. Client truy vấn Namenode để đưa thông tin về thao tác đọc. Namenode kết nối song song tới các datanode để lấy dữ liệu, sau đó trả về cho client.
D. Client truy vấn Namenode để biết được vi trí các chunks. Nếu Namenode không biết về vị trí các chunk thì namenode sẽ hỏi các datanode. Sau đó Namenode gửi lại thông tin vị trí các chunk cho client. Client kết nối song song tới các datanode để đock các chunk.

Câu 18: Mô tả cách thức một client ghi dữ liệu trên HDFS.
A. Client kết nối tới Namenode chỉ định muốn ghi vào chunk nào. Namenode trả về vị trí các chunk cho client. Client ghi đồng thời vào các datanode.
B. Client kết nối tới Namenode chỉ định khối lượng dữ liệu cần ghi. Namenode trả về vị trí các chunk cho client. Client ghi chunk tới datanode đầu tiên, sau đó các datanode tự động thực thi nhân bản. Quá trình ghi kết thúc thi tất cả các chunk và các nhân bản đã được ghi thành công.
C. Client kết nối tới Namenode chỉ định khối lượng dữ liệu cần ghi. Namenode trả về vị trí các chunk cho client. Client ghi đồng thời các chunk vào datanode. Với mỗi chunk, các datanode thực thi nhân bản tự động sau khi thao tác ghi thành công.

Câu 19: Cơ chế chịu lỗi của datanode trong HDFS?
A. Sử dụng Zookeeper để quản lý các thành viên datanode trong cụm.
B. Sử dụng cơ chế heartbeat, định kỳ các datanode thông báo về trạng thái cho Namenode.
C. Sử dụng cơ chế heartbeat, Namenode định kỳ hỏi các datanode về trạng thái tồn tại của các datanode.

Câu 20: Cơ chế tổ chức dữ liệu của Datanode trong HDFS?
A. Các chunk là các tệp tin trong hệ thống tệp tin cục bộ của máy chủ datanode.
B. Các chunk là các vùng dữ liệu liên tục trên ổ cứng của máy chủ datanode.
C. Các chunk được lưu trữ tin cậy trên datanode theo cơ chế RAID.

Câu 21: Cơ chế nhân bản dữ liệu trong HDFS?
A. Namenode quyết định vị trí các nhân bản của các chunk trên datanode.
B. Datanode là primary quyết định vị trí các nhân bản của các chunk tại các secondary datanode.
C. Client quyết định vị trí lưu trữ các nhân bản với từng chunk.

Câu 22: HDFS giải quyết bài toán một điểm hỏng hóc duy nhất (single-point-of-failure) cho Namenode bằng cách nào?
A. Sử dụng thêm secondary namenode theo cơ chế active-active. Cả Namenode và Secondary namenode cùng online trong hệ thống
B. Sử dụng Secondary namenode theo cơ chế active-passive. Secondary namenode chỉ hoạt động khi có vấn đề với Namenode.

Câu 23: CSDL nào dưới đây không phải là NoSQL
A. Microsoft SQL server
B. MongoDB

C. Cassandra
D. Không phải các đáp án đã đưa ra

Câu 24: Chọn phát biểu đúng khi nói về MongoDB
A. Các văn bản có thể chứa nhiều cặp key-value hoặc key-array, hoặc các văn bản lồng (nested documents).
B. MongoDB có các trình điều khiển driver cho nhiều ngôn ngữ lập trình khác nhau.
C. MongoDB hay các NoSQL có khả năng khả mở tốt hơn các CSDL quan hệ truyền thống.
D. Tất cả các phương án đã đưa ra.

Câu 25: Đâu là một dạng của NoSQL
A. MySQL
B. JSON
C. Key-value store
D. OLAP

Câu 26: Đâu là một CSDL dạng cột mở rộng?
A. Cassandra
B. Riak (key-value)
C. MongoDB (document-based)
D. Redis (key-value)

Câu 27: Chọn phát biểu sai
A. NoSQL yêu cầu lược đồ CSDL phải được định nghĩa trước khi thêm dữ liệu
B. NoSQL cho phép thêm vào dữ liệu mà không cần định nghĩa trước lược đồ dữ liệu
C. NoSQL được đưa ra nhằm bổ sung các giải pháp mà CSDL truyền thống không đáp ứng tốt

Câu 28: Cơ chế mà NoSQL sử dụng để tăng khả năng chịu lỗi
A. Phân mảnh và phân tán dữ liệu ra nhiều máy chủ
B. Nhân bản (Replication)
C. Giao diện truy vấn đơn giản hơn so với CSDL quan hệ truyền thống

Câu 29: CSDL nào phù hợp với dữ liệu mạng xã hội, dữ liệu có sự liên kết
A. Key-value
B. Document store
C. Graph store
D. Columnar store

Câu 30: Chọn phát biểu đúng về NoSQL
A. Không hỗ trợ các truy vấn SQL
B. Không thể được sử dụng kết hợp với các CSDL quan hệ
C. Rất phù hợp cho các tập dữ liệu phân tán quy mô lớn
D. Đáp ứng khả năng xử lý giao dịch với tính nhất quán chặt

Câu 31: Đâu không phải là tính năng mà NoSQL nào cũng đáp ứng
A. Phù hợp với dữ liệu lớn
B. Khả năng mở rộng linh hoạt
C. Tính sẵn sàng cao

Câu 32: Tình huống triển khai nào phù hợp với NoSQL
A. Khi cần đáp ứng về tính toàn vẹn của dữ liệu (data integrity)

B. Khi cần đáp ứng cao về vấn đề bảo mật dữ liệu
C. Khi cần lưu trữ hiệu quả dữ liệu lớn
D. Khi lược đồ dữ liệu không quá phức tạp Ko chọn D

Câu 33: NoSQL có đặc điểm nào dưới đây?
A. Mở rộng theo chiều ngang, tinh chỉnh được tính sẵn sàng của hệ thống
B. Không thể sử dụng SQL để truy vấn dữ liệu NoSQL
C. Mở rộng theo chiều dọc, thiết kế đơn giản, khó tinh chỉnh tính sẵn sàng của hệ thống
D. Mở rộng theo chiều dọc, thiết kế phức tạp, tinh chỉnh được tính sẵn sàng của hệ thống

Câu 34: Công ty nào đã phát triển Apache Cassandra giai đoạn đầu tiên?
A. Facebook
B. Google
C. Linkedin
D. Twitter

Câu 35: Hệ thống nào cho phép đọc ghi dữ liệu tại vị trí ngẫu nhiên, thời gian thực tới hàng terabyte dữ liệu
A. Hbase
B. Flume
C. Pig
D. HDFS

Flume không phải là hệ thống lưu trữ dữ liệu mà là một công cụ để chuyển dữ liệu
Pig không phải là hệ thống lưu trữ dữ liệu mà là một công cụ để phân tích dữ liệu
HDFS được tối ưu hóa cho việc xử lý theo lô (batch processing), không phải cho các thao tác thời gian thực.

Câu 36: Phát biểu nào sai về Hbase
A. Hbase có lệ thuộc vào các dịch vụ cung cấp bởi Zookeeper
B. Hbase có hệ thuộc vào các dịch vụ cung cấp bởi HDFS
C. Hbase không hỗ trợ versioning
D. Hbase hỗ trợ truy vấn dạng SQL

Câu 37: Thao tác nào không được hỗ trợ bởi Hbase
A. Put
B. Get
C. Scan
D. Multiput
E. Join

Câu 38: Hbase có thể được sử dụng cho kiểu dữ liệu nào
A. Dữ liệu có cấu trúc
B. Dữ liệu phi cấu trúc
C. Dữ liệu bán cấu trúc
D. Tất cả các phương án được đưa ra.

Câu 39: Điều gì xảy ra nếu chúng ta chọn Hbase row key là timestamp tại thời điểm insert dữ liệu?
A. Insert sẽ nhanh hơn so với row key là dữ liệu khác
B. Insert sẽ chậm hơn so với row key là dữ liệu khác
C. Tùy trường hợp

Việc sử dụng timestamp làm row key có thể dẫn đến hiện tượng hot spot (điểm nóng) trong HBase. Hot spot xảy ra khi các dữ liệu mới được insert vào bảng HBase đồng thời với timestamp giống nhau hoặc gần nhau, khiến cho các dòng dữ liệu mới đều được lưu trữ trên cùng một Region Server.

Câu 40: Dữ liệu trả về từ Hbase luôn được sắp xếp theo trật tự nào?
A. Rowkey, column family, column qualifier, timestamp
B. Timestamp, column qualifier, column family, row key
C. Row key, column family, column qualifier

Câu 41: Phát biểu nào sai về Hfile trong Hbase?
A. Hfile chứa một tập hợp các dòng bản ghi trong Hbase table
B. Nhiều Hfile có thể được gộp lại thành 1 Hfile lớn theo những khoảng thời gian nhất định
C. Một version của 1 dòng hay 1 bản ghi trong Hbase table có thể được phân rã trên nhiều Hfile khác nhau
D. Nhiều Hfile có thể được gộp lại thành 1 Hfile lớn khi cần thiết

Câu 42: Các đặc điểm của virtual node trên AmazonDB. Chọn phương án sai
A. Mỗi node vật lý có thể được ánh xạ thành nhiều node ảo, nằm liên tiếp nhau trong vòng tròn không gian khoá.
B. Số lượng các node ảo đối với mỗi node vật lý là khác nhau tùy vào từng node vật lý.
C. Số lượng các node ảo bắt buộc cần phải căn cứ vào khả năng lưu trữ của node vật lý.
D. Node ảo đóng vai trò quan trọng trong bài toán cân bằng tải và hiệu năng khi một node vật lý ra hoặc kết nối vào cụm.

Câu 43: Phát biểu nào đúng về Amazon DynamoDB
A. DynamoDB là zero-hop DHT
B. DynamoDB là one-hop DHT
C. DynamoDB là multiple-hop DHT

Câu 44: Phát biểu nào đúng về Quorum trong Amazon DynamoDB
A. Với N là tổng số nhân bản, R là số nhân bản cần đọc trong 1 thao tác đọc. W là số nhân bản cần ghi trong 1 thao tác ghi. N = R + W
B. Với N là tổng số nhân bản, R là số nhân bản cần đọc trong 1 thao tác đọc. W là số nhân bản cần ghi trong 1 thao tác ghi. N > R + W
C. Với N là tổng số nhân bản, R là số nhân bản cần đọc trong 1 thao tác đọc. W là số nhân bản cần ghi trong 1 thao tác ghi. N < R + W

Câu 45: Phát biểu nào sai về Amazon DynamoDB
A. Dynamo có cơ chế cho phép 1 node đứng ra thực hiện thao tác ghi cho 1 node đang bị rớt khỏi mạng tạm thời và trả lại bản ghi này khi node đó quay lại hệ thống.
B. Dynamo node sử dụng merkle tree để nhanh chóng kiểm tra sự không nhất quán của các nhân bản dữ liệu
C. Dynamo có cơ chế versioning sử dụng vector clock
D. Trong trường hợp có xung đột giữa các phiên bản dữ liệu, sự xung đột được giải quyết tại thao tác ghi.

Câu 46: Phát biểu nào sai về Presto?
A. Presto là một engine truy vấn SQL hiệu năng cao, phân tán cho dữ liệu lớn
B. Presto cho phép tích hợp với các công cụ Business Intelligence
C. Presto được quản lý bởi Apache Software foundation
D. Presto được quản lý bởi Presto Software foundation

Câu 47: Phát biểu nào sai về Presto?
A. Presto có thể truy vấn nhiều data storages khác nhau như HDFS, Cassandra

B. Presto không truy vấn được dữ liệu trong MySQL, MS SQL và các CSDL quan hệ truyền thống
C. Presto thường nhanh hơn Hive hay Pig

Câu 48: Phát biểu nào sai về Presto?
A. Lược đồ thực thi phân tán (distributed query plan) gồm nhiều màn (stages)
B. Stage là phần công việc có thể được thực thi song song bởi nhiều workers
C. 1 tác vụ (task) là một đơn vị công việc được gán cho một worker ứng với 1 stage
D. 1 tác vụ gồm nhiều pipelines, mỗi pipelines là 1 chuỗi các thao tác
E. 1 tác vụ xử lý một split dữ liệu Mỗi task thường xử lý một hoặc nhiều split dữ liệu

Trong Presto, 1 tác vụ (task) không xử lý một split dữ liệu mà thực hiện xử lý trên một phần dữ liệu của các bảng hoặc bộ dữ liệu được chia thành các phần (partition). Các split dữ liệu là các phần nhỏ hơn của partition và được gửi đến các worker để xử lý song song. Mỗi tác vụ có thể chứa nhiều split dữ liệu để thực hiện trên nhiều worker.

Câu 49: Phát biểu nào sai về cơ chế scheduling của Presto?
A. Một task có thể được lập lịch chạy trên bất kỳ worker nào
B. Stage có thể được lập lịch all-at-once
C. Stage có thể được lập lịch theo giai đoạn
D. Split được gán cho task theo cơ chế lazy

Câu 50: Phát biểu nào đúng về Presto?
A. Các stage được thực thi theo cơ chế pipeline, không có thời gian chờ giữa các stage như Map Reduce
B. Presto có cơ chế chịu lỗi khi thực thi truy vấn vì là all-at-once nên nếu query fail -> all fail
C. Presto cho phép xử lý kết tập dữ liệu mà kích thước lớn hơn kích thước bộ nhớ trong
không cho phép kết tập dữ liệu nếu kích thước ko vừa kích thước bộ nhớ trong => vì thế query fail

Câu 51: Phát biểu nào sau sai về Kafka?
A. Kafka quản lý các luồng thông điệp (messages) thành các nhóm gọi là các Topics.
B. Tiến trình quảng bá message lên cụm Kafka gọi là publishers.
C. Tiến trình đăng ký theo dõi các topics gọi là consumers
D. Các máy chủ chạy Kafka gọi là các brokers.

Trong Kafka, tiến trình quảng bá (publish) message lên cụm Kafka được gọi là producers (hoặc Kafka producers). Các producers gửi các message vào các topic trong Kafka.

Câu 52: Phát biểu nào sau đây sai về Kafka
Kafka chỉ đảm bảo thứ tự của các message trong một partition, không phải trên toàn bộ topic
A. Các topic gồm nhiều partition
B. Partition được nhân bản ra nhiều brokers.
C. Kafka bảo đảm thứ tự của các message với mỗi topics.
D. Message sau khi được tiêu thụ (consume) thì không bị xoá Các message được lưu lại theo chính sách retention.

Câu 53: Phát biểu nào sau đây sai về Kafka?
A. Mỗi partition có 1 leader và nhiều followers. ko hoặc nhiều follower
B. Tất cả các thao tác ghi, đọc được xử lý bởi leader, follower làm theo leader.
C. Nếu leader bị lỗi, 1 follower sẽ thay thế trở thành leader mới

Câu 54: Phát biểu sau đây đúng hay sai: Trong cụm Kafka, 1 server đóng vai trò leader, các server còn lại đóng vai trò follower.
A. Đúng

B. Sai

Câu 55: Phát biểu nào sai về Kafka?
A. Các message trên Kafka được lưu lại theo thời gian (time-based)
B. Các message trên Kafka được lưu lại theo kích thước partition (size-based)
C. Các message trên Kafka được lưu lại trước khi thực hiện compaction
D. Message sau khi được tiêu thụ bởi tất cả các consumer thì bị xoá.

theo cấu hình retention -> nên ko chắc là sẽ luôn bị xóa

Câu 56: Phát biểu nào sai về Kafka?
A. Kafka producer quyết định message sẽ được gửi đến partition nào trong topic.
B. Thứ tự của message trong mỗi partition do key của message quyết định.
C. Kafka producer có thể gửi message đến nhiều broker khác nhau.

Trong Kafka, thứ tự của các message trong mỗi partition được bảo đảm duy nhất và bất biến (immutable) dựa trên thời gian mà chúng được gửi đến Kafka.

Câu 57: Phát biểu nào sau đây sai về Kafka?
A. Nhiều consumer có thể cùng đọc 1 topic.
B. 1 message chỉ có thể được đọc bởi 1 consumer trong 1 consumer group.
C. 1 message có thể được đọc bởi nhiều consumer khác nhau.
D. Số lượng consumer phải ít hơn hoặc bằng số lượng partitions.

Một message có thể được đọc bởi nhiều consumer nếu chúng thuộc các consumer groups khác nhau.

Nếu số lượng consumers vượt quá số lượng partitions-> Consumer bị thừa đấy vẫn ở trạng thái chờ; có thể được sử dụng để chuyển đổi dự phòng

Câu 58: Đâu là vấn đề khi xử lý dữ liệu lớn với MapReduce?
A. Xử lý luồng dữ liệu lớn
B. Xử lý dữ liệu lớn trong thời gian tương tác
C. Xử lý chuỗi các công việc
D. Xử lý dữ liệu lớn theo lô (Bulk processing)

Câu 59: Đâu là ưu điểm của Spark so với MapReduce?
A. Hỗ trợ tốt cho xử lý chuỗi các biến đổi
B. Có thể khai phá dữ liệu trong thời gian tương tác
C. Khai thác bộ nhớ trong thay vì sử dụng hệ thống lưu trữ ngoài như HDFS
D. Có khả năng chịu lỗi

vì cả 2 đều chịu lỗi

Câu 60: Đâu là cơ chế chịu lỗi của Apache Spark?
A. Chịu lỗi qua cơ chế huyết thống
B. Chịu lỗi qua cơ chế nhân bản
C. Chịu lỗi qua cơ chế lưu lại lịch sử nhiều phiên bản

Spark không sử dụng cơ chế nhân bản dữ liệu như HDFS

Spark không lưu trữ nhiều phiên bản của dữ liệu, mà chỉ lưu lịch sử các biến đổi (lineage) để tái tạo dữ liệu khi cần.

Khi một phần tử trong quá trình tính toán bị lỗi (ví dụ: một máy tính bị hỏng), Spark có khả năng tự động khôi phục lại công việc mà phần tử đó đang thực hiện bằng cách sử dụng dữ liệu được lưu trữ có sẵn (ví dụ: dữ liệu đã được lưu trữ trên HDFS hoặc các hệ thống lưu trữ phân tán khác). Quá trình khôi phục này đảm bảo rằng việc tính toán không bị mất hoàn toàn và tiếp tục chạy mà không phải thực hiện lại từ đầu.

Câu 61: Đâu là đặc điểm của RDD (Resilient distributed dataset) của Spark?
A. Được chia thành các phân mảnh (partition)
B. Người lập trình có thể quyết định số các phân mảnh của mỗi RDD
C. Người sử dụng không thể quyết định số các phân mảnh của mỗi RDD
D. Có khả năng chịu lỗi

huyết thống

Câu 62: Đâu là đặc điểm của RDD (Resilient distributed dataset) của Spark
A. Được thiết kế để tối ưu cho các biến đổi thô, theo lô
B. Được thiết kế hỗ trợ các cập nhật đơn lẻ tới mức từng bản ghi
C. Có khả năng tự động tái tạo lại khi bị lỗi qua cơ chế nhân bản

Câu 63: Đâu là các thao tác có thể thực hiện trên RDD (Resilient distributed dataset) của Spark?
A. Thực hiện các biến đổi (transformation)
B. Thực hiện các hành động (action)
C. Yêu cầu Spark lưu RDD ở bộ nhớ đệm
D. Thực hiện các biến đổi mà xoá các bản ghi trong RDD
E. Thực hiện các biến đổi mà cập nhật các bản ghi trong RDD

Câu 64: Các biến đổi (transformation) trên Spark có đặc điểm gì?
A. Thực hiện theo cơ chế lười biếng, khi nào một hành động (action) cần tới phép biến đổi trước đó phải thực hiện thì mới phải thực hiện
B. Mỗi phép biến đổi trên RDD được thực thi bởi một hay nhiều Spark worker
C. Các biến đổi (transformation) luôn tạo ra RDD mới có cùng số partition với RDD đầu vào

Câu 65: Đâu là đặc điểm của Spark Streaming?
A. Spark Streaming rời rạc hóa luồng dữ liệu đầu vào thành DStream là chuỗi liên tục của các RDD nhỏ
B. Spark streaming xử lý liên tục từng bản ghi ngay khi nhận được từ luồng dữ liệu đầu vào

Câu 66: Đâu là đặc điểm của Spark streaming?
A. Có thể nhận đầu vào là các luồng dữ liệu từ Kafka
B. Có thể nhận đầu vào là các tệp tin trên HDFS
C. Không thể thực hiện các truy vấn SQL

Câu 67: Spark structured streaming có đặc điểm gì?
A. Vẫn xử lý luồng dữ liệu như là chuỗi các lô RDD nhỏ
B. Về mặt logic, coi luồng như một bảng dữ liệu liên tục tăng thêm các bản ghi
C. Định kỳ, truy vấn trên luồng chỉ trả ra kết quả của việc thực hiện truy vấn cho trên các bản ghi mới xuất hiện

Câu 68: Kiến trúc xử lý dữ liệu Lambda có đặc điểm gì?
A. Kết hợp xử lý dữ liệu theo lô và theo luồng
B. Giúp giải quyết vấn đề độ trễ từ khi dữ liệu được thập tới kết quả phân tích của mô hình xử lý theo lô
C. Giúp giải quyết vấn đề nhược điểm của xử lý theo luồng là kết quả phân tích không khai thác được toàn bộ dữ liệu trong lịch sử.
D. Có kiến trúc gồm 2 tầng: tầng xử lý theo lô và tầng xử lý theo luồn
E. Bao gồm các tiến trình ETL (extract, transform, load) đưa dữ liệu vào hồ dữ liệu (data lake)