1. Cau 1
2. Đâu là vấn đề khi xử lý dữ liệu lớn với MapReduce?
    a. Xử lý dữ liệu lớn theo lô (Bulk processing)
    b. Xử lý chuỗi các công việc
    c. Xử lý luồng dữ liệu lớn
    d. Xử lý dữ liệu lớn trong thời gian tương tác.

3. Ưu điểm của kiến trúc SAN (Storage area network)?
    a. Quản trị dễ dàng hơn so với NAS.
    b. Máy khách có thể kết nối tới SAN bằng đường truyền Ethenet thông thường (Chuẩn kết nối TCP/IP).
    c. Hiệu năng, băng thông tốt hơn với NAS.

4. Thao tác nào không được hỗ trợ bởi Hbase?
    a. Join
    b. Put
    c. Scan
    d. Multiput
    e. Get
5. Thế nào là UNIX semantic?
    a. Tệp tin là chỉ đọc, không cho phép cập nhật và ghi đè. Mọi tiến trình đều có thể đọc tệp tin đồng thời.
    b. 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.
    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.

6. Tình huống triển khai nào phù hợp nới NoSql?
        a. Khi lược đồ dữ liệu không quá phức tạp.
        b. Khi cần lưu trữ hiệu quả dữ liệu lớn.
        c. Khi cần đáp ứng về tính toàn vẹn của dữ liệu (data integrity).
        d. Khi cần đáp ứng cao về vấn đề bảo mật dữ liệu.
7. Chọn phát biểu đúng về NoSql?
        a. Không thể được sử dụng kết hợp với các CSDL quan hệ.
        b. Rất phù hợp cho các tập dữ liệu phân tán quy mô lớn.
        c. Đáp ứng khả năng xử lý giao dịch với tính nhất quán chặt.
        d. Không hỗ trợ truy vấn SQL.
8. Phát biểu nào sau đây sai về Kafka?
        a. Các topic gồm nhiều partion.
        b. Message sau khi được tiêu thụ (consume) thì không bị xóa.
        c. Kafka bảo đảm thứ tự của message với mỗi topics.

d. Partion được nhân bản ra nhiều brokers.

9. Đâu là đặc điểm của spark streaming?
    a. Spark streaming xử lý liên tục từng bản ghi ngay khi nhận được luồng dữ liệu đầu vào.
    b. Spark streaming rời rạc hóa luồng dữ liệu đầu vào thành Dstream là chuỗi liên tục cảu các RDD nhỏ.
10. Đâu là 1 dạng của NoSQL?
    a. MySQL.
    b. OLAP.
    c. JSON.
    d. Key-value store.
11. Mô tả cách thức một client đọc dữ liệu trên HDFS.
    a. Client truy vấn Namenode để biết được vị trí các chunks. Clients kết nối song song tới các datanode để đọc các chuck.
    b. Client truy vấn Namenode để biết được vị trí các chucks. Nếu Namenode không biết về vị trí các chuck thì namenode sẽ hỏi các datanode. Sau đó Namenode gửi lại thông tin vị trí các chuck cho client. Client kết nối song song tới các datanode để dọc các chuck.
    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 thông báo tới namenode để bắt đầu quá trình đọc sau đó client truy vấn các datanode để trực tiếp đọc các chunks.
12. 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. Sai.
    b. Đúng.
13. Phát biểu nào sau đây sai về Kafka?
    a. Mỗi partion có 1 leader về nhiều followers.
    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.
14. 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.
15. Vai trò của YARN?
    a. 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.
    b. Quản lý và phân phối tài nguyên trong cụm Hadoop.
    c. Cung cấp giao diện người dùng mức cao, biến đổi truy vấn thành các job Mapreduce.
16. Phát biểu nào sai về Hfile trong Hbase?
    a. Một version của 1 dòng hay 1 bản ghi trong Hbase table có thể được phân rã trên nhiều Hfile khác nhau.
    b. Nhiều Hfile có thể được gộp lại thành 1 Hfile lớn theo những khoảng thời gian nhất định.

c. Nhiều Hfile có thể được gộp lại thành 1 Hfile lớn khi cần thiết.
d. Hfile chứa một tập hợp các dòng bản ghi trong Hbase table.

17. Phát biểu nào sau đây sai về Kafka?
    a. 1 message có thể được đọc bởi nhiều consumer khác nhau.
    b. 1 message chỉ có thể được đọc bởi 1 consumer trong 1 consumer group.
    c. Số lượng consumer phải ít hơn hoặc bằng số lượng partions.
    d. Nhiều consumer có thể cùng đọc 1 topic.
18. Đâu là cơ chế chịu lỗi của Apache Spark?
    a. Chịu lỗi qua cơ chế nhân bản.
    b. Chịu lỗi qua cơ chế lưu lại lịch sử nhiều phiên bản.
    c. Chịu lỗi qua cơ chế huyết thống.
19. 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.
20. Ưu điểm cảu kiến trúc NAS (Network attached Storage)?
    a. Máy khách có thể kết nối tới NAS bằng đường truyền Ethernet thông thường (chuẩn kết nối TCP/IP).
    b. Đơn giản hóa việc chia sẻ dữ liệu.
    c. Tính khả mở cao.
21. Cơ chế mà NoSQL sử dụng để tăng khả năng chịu lỗi.
    a. Giao diện truy vấn đơn giản hốn với CSDL quan hệ truyền thống.
    b. Nhân bản (Replication).
    c. Phân mảnh và phân tán dữ liệu ra nhiều máy chủ.
22. Phát biểu nào sai về Presto?
    a. Presto được quản lý bởi Presto Software foundation.
    b. Presto cho phép tích hợp với các công cụ Business Intelligence.
    c. Presto được quản lý bởi Apache Software foundation.
    d. Presto là một engine truy vấn SQL hiệu năng cao, phân tán cho dữ liệu lớn.
23. Cơ chế tổ chức dữ liệu của Datanode trong HDFS?
    a. Các chunk là các vùng dữ liệu liên tục trên ổ cứng của máy chủ datanode.
    b. Các chunk được luê trữ tin cậy trên datanode theo cơ chế RAID.
    c. Các chunk là các tệp tin trong hệ thống tệp tin cục bộ của máy chủ datanode.
24. 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 lưu trữ tin cậy trên ổ cứng theo cơ chế RAID.
    c. Các công việc cần tính toán được phân mảnh thành các tác vụ độc lập.
    d. Các tệp tin được phân mảnh, các mảnh được nhân bản ra các node khác trên cụm.
25. Phát biểu nào sai về Presto?
    a. Presto thường nhanh hơn Hive hay Pig.
    b. Presto không truy vấn được dữ liệu trong MySQL, MS SQL và các CSDL quan hệ truyền thống.
    c. Presto có thể truy vấn nhiều data storages khác nhau như HDFS, Cassandra.

26. Chọn phát biểu sai.
    a. NoSQL được đưa ra nhằm bổ sung các giải pháp mà CSDL truyền thống không đáp ứng tốt.
    b. NoSQL cho phép thêm vào dữ liệu mà không cần định nghĩa trước lược đồ dữ liệu.
    c. NoSQL yêu cầu lược đồ CSDL phải được định nghĩa trước khi thêm dữ liệu.
27. Đâu là ưu điểm của Spark so với MapReduce?
    a. Hỗ trợ tốt cho xử lý chuỗi các biến đổi.
    b. Có khả năng chịu lỗi.
    c. Có thể khai phá dữ liệu trong thời gian tương tác.
    d. Khai thác bộ nhớ trong thay vì sử dụng hệ thống lưu trữ ngoài như HDFS.
28. Kiến trúc xử lý dữ liệu Lambda có đặc điểm gì?
    a. 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.
    b. Bao gồm các tiến trình ETL (extract, transform, load) đưa dữ liệu vào hồ dữ liệu (data table).
    c. Có kiến trúc gồm 2 tầng: tầng xử lý theo lô và tầng xử lý theo luồng.
    d. 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.
    e. Kết quả xử lý dữ liệu theo lô và theo luồng.
29. Các đặc điểm của HDFS. Chọn đáp án sai.
    a. Hỗ trợ cơ chế phân quyền và kiểm soát người dùng của UNIX.
    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. Tối ưu cho các tệp tin có kích thước lớn.
30. Điều gì xảy ra nếu chúng ta chọn Hbase row key là timestamp tại thời điểm insert dữ liệu?
    a. Insert sẽ chậm hơn so với row key là dữ liệu khác.
    b. Tùy trường hợp.
    c. Insert sẽ nhanh hơn so với row key là dữ liệu khác.
31. Các đặc điểm của virtual node trên AmazonDB. Chọn phương án sai.
    a. 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.
    b. Số lượng các node ảo đối với mỗi node vật lý là khác nhau tùy vào từng node vật lý.
    c. Mỗi node vật lý có thể dược ánh xạ thành nhiều node ảo, nằm liên tiếp nhau trong vòng tròn không gian khóa.
    d. Số lượng các node ảo bắt buộc cần phải căn cứ vào khả năng lưu trữ của node vật lý.
32. NoSQL có đặc điểm nào dưới đây?
    a. Mở rộng theo chiều dọc, thiết kế phức tạp, tinh chỉnh được tính sẵn sàng của hệ thống.
    b. Mở rộng theo chiều ngang, tinh chỉnh được tính sẵn sàng của hệ thống.
    c. Không thể sử dụng SQL để truy vấn dữ liệu NoSQL.
    d. Mở rộng theo chiều dọc, thiết kế đơn giản, khó tinh chỉnh tính sẵn sàng của hệ thống.
33. Các mục tiêu chính của Apache Hadoop?
    a. Xử lý dữ liệu lớn mạnh mẽ.
    b. Lưu trữ dữ liệu khả mở và Xử lý dữ liệu lớn mạnh mẽ.
    c. Lưu trữ dữ liệu khả mở.

d. Lưu trữ dữ liệu khả mở , xử lý dữ liệu lớn mạnh mẽ và trực quan hóa dữ liệu hiệu quả.
e. Trực quan hóa dữ liệu hiệu quả.

34. Thành phần nào không thuộc thành phần lõi của Hadoop?
    a. Apache Hbase.
    b. Mapreduce framework.
    c. Apache Zookeeper.
    d. YARN: yet another resource negotiator.
    e. Hệ thống tập tin phân tán HDFS.
35. Đâu là kỹ thuật có thể được dùng để thích nghi các giải thuật học máy cho dữ liệu lớn?
    a. Tất cả các ý (1),(2),(3).
    b. Các ý (2) và (3)
    c. (2) song song hóa trên Mapreduce hay Spark.
    d. (1) Sub-sampling, principal component analysis, feature extraction và feature selection.
    e. (3) các kiến trúc mới xử lý luồng liên tục như mini-batch, complex event processing.
36. Phát biểu nào đúng về Presto?
    a. Presto cho phép xử lý kết tập dữ liệu mà kích thước lớn hơn kích thước bộ nhớ trong.
    b. Presto có cơ chế chịu lỗi khi thực thi truy vấn.
    c. Các stage được thực thi theo cơ chế pipeline, không có thời gian chờ giữa các stage như Map Reduce.
37. Công cụ nào có thể sử dụng để hỗ trợ import, export dữ liệu vào ra hệ sinh thái Hadoop?
    a. Oozie.
    b. Flume.
    c. Hive.
    d. Sqoop.
38. Đâu là các thao tác có thể thực hiện trên RDD (Resilient distributed dataset) của Spark?
    a. Thực hiện các biến đổi mà xóa các bản ghi trong RDD.
    b. Thực hiện các biến đổi mà cập nhật các bản ghi trong RDD.
    c. Yêu cầu Spark lưu RDD ở bộ nhớ đệm.
    d. Thực hiện các biến đổi (transformation).
    e. Thực hiện các hành động (action).
39. Các biến đổi (transformation) trên Spark có đặc điểm gì?
    a. Mỗi phép biến đổi trên RDD được thực thi bởi một hay nhiều Spark worker.
    b. 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.
    c. Các biến đổi (tranformation) luôn tạo ra RDD mới có cùng số partion với RDD đầu vào.
40. Ưu điểm của hệ thống tệp tin phân tán là gì?
    a. Cho phép người dùng có cái nhìn hợp nhất(như nhau) về toàn bộ dữ liệu trong hệ thống.
    b. Đơn giản hóa việc chia sẻ dữ liệu.
    c. Tập trung hóa việc quản trị.