1. Spark hỗ trợ Cluster Manager :
    A. MESOS
    B. YARN
    C. standalone cluster manager
    D. Không có đáp án đúng
2. Output operation khi thao tác với Dstream :
    A. saveAsTextFiles,
    B. saveAsHadoopFiles,
    C. reduceByKeyAndWindow
3. Đáp án nào ko phải "Transformation" khi thao tác với Dstream :
    A. reduceByWindow
    B. window
    C. foreachWindow
    D. countByWindow
4. Bản chất của Dstream là gì?
    A. Là một chuỗi liên tục RDD
    B. Là 1 chuỗi liên tục DataFrame
    C. Là 1 chuỗi liên tục DataSet

5. Trong hệ sinh thái của Spark không có công cụ hay thành phần nào sau đây :
    A. Mlib
    B. GraphX
    C. Sqoop
    D. ClusterManagers

6. Các mục tiêu chính của Apache Hadoop :
    A. Lưu trữ dữ liệu khả mở
    B. Xử lý dữ liệu mạnh mẽ
    C. Trực quan hóa 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 hóa dữ liệu hiệu quả

7. Phát biểu nào sau đây **không đúng** về Apache Hadoop
    A. Xử lý dữ liệu phân tán với mô hình lập trình đơn giản

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

8. 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

9. 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 toán hoặc là node lưu trữ dữ liệu
   C. Các node tham gia vào cụm đóng cả 2 vài trò tính toán và lưu trữ
   D. Các node tham gia vào cụm có thể có cấu hình, độ tin cậy cao

10. 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 đc lưu trữu 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

11. 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 chuck (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

12. 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 sẽ hỏi các datanode. Sau đó 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 đọc các chunk

13. Mô tả cách thức 1 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 khi 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.

14. 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 kì các datanode thông báo về trạng thái cho Namdnode
C. Sử dụng cơ chế heartbeat, Namenode định kì hỏi các datanode về trạng thái của datanode.

15. 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

16. Cơ chế nhân bản dữ liệu trong HDFS
A. Namenode quyết định vị trí các nhân bản của các chunk trên datanode.
B. Namenode 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.

17. HDFS giải quyết bài toán 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.

18. Đầu vào dữ liệu cho chương trình Spark có thể là :
A. Local file
B. HDFS, NFS
C. Amazon S3, Elasticsearch
D. Cả 3 phương án trên

19. Đâu là lệnh lưu dữ liệu ra ngoài chương trình Spark

A. Input.saveAsTextFile('file://usr/zeppelin/notebook/dataset/new.txt')
B. Input.saveAsTextFile('/usr/zeppelin/notebook/dataset/new.txt')
C. Input.saveAs ('file:///usr/zeppelin/notebook/dataset/new.txt')
D. Input.saveAsTextFile:'file:///usr/zeppelin/notebook/dataset/new.txt'

20. Đâu là cách submit đúng 1 job lên Spark cluster hoặc chế độ local :

A. ./spark-submit wordcount.py README.md
B. ./spark-submit README.md wordcount.py
C. Spark-submit README.md wordcount.py
D. Phương án A và C

21. Câu lệnh MapReduce trong Spark dưới đây, chia mỗi dòng thành từ dựa vào delimiter nào.

Input.flatMap(lambda x: x.split('\t')).map(lamda x: (x,1)).reduceByKey(add)

A. Tab
B. Dấy cách
C. Dấy hai chấm
D. Dấu phẩy

22. Data Pipeline nào sau đây là đúng trên Spark

A. Spark -> RabbitMQ -> Elasticsearch -> Hiển thị
B. Dữ liệu sensor -> RabbitMQ -> Elasticsearch -> Spark -> hiển thị
C. Dữ liệu sensor -> Elasticsearch -> RabbitMQ -> Spark -> Hiển thị
D. Spark -> Elasticsearch -> Hiển thị

23. Spark có thể chạy ở chế độ nào khi chạy trên nhiều máy?

A. Chạy trên YARN
B. Chạy trên Zookeeper
C. Phương án A và B đều sai
D. Phương án A và B đều đúng

24. Mục đích của sử dụng RabbitMQ là gì?
    A. Lưu trữ dữ liệu
    B. Tránh dữ liệu bị mất mát
    C. Hiển thị dữ liệu
    D. Phân tích dữ liệu

25. Mục đích của sử dụng Spark ML là gì ?
    A. Chạy MapReduce
    B. Chạy các thuật toán dự đoán
    C. Tính toán phân tán
    D. Cả B và C

26. Mục đích của lệnh sau đây là gì?
    (trainingData, testData) = dataset.randomSplit([0.8, 0.2], seed = 10)
    A. Chia dữ liệu học và dữ liệu kiểm tra
    B. Chạy chương trình học
    C. Tạo dữ liệu ngẫu nhiên cho dữ liệu học và kiểm tra
    D. Chạy chương trình dự đoán

27. Label và Feature của câu lệnh bên dưới có ý nghĩa j ?
    LogisticRegression(labelCol = “label”, featuresCol = “features”,maxIter = 10)
    A. Dữ liệu đầu vào được gán là feature và dự đoán được gán vào label
    B. Dữ liệu đầu vào được gán là label và kết quả của dữ liệu đầu vào được gán là feature
    C. Dữ liệu đầu vào được gán là feature và kết quả của dữ liệu đầu vào đó được gán là label
    D. Dữ liệu đầu vào được gán là label và kết quả dự đoán được gán vào feature