Module 08 & 09

Câu 1: 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

Câu 2: 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'

Câu 3: Đâu là cách submit đúng một 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

Câu 4: Câu lệnh MapReduce trong Spark dưới đây, chia mỗi dòng thành từ dựa vào delimiter nào.

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

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

Module 12&13

Câu 5: 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 → Elasticserach→RabbitMQ→Spark→Hiển thị
D. Spark→Elasticsearch→Hiển thị

Câu 6: 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

Câu 7: 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. Cả 2 phương án A và B

Module 15 & 16

Câu 8: 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 and C

Câu 9: Mục đích của lệnh sau đây là gì?

(trainingData, testData) = dataset.randomSplit([0.8, 0.2], seed=100)

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

Câu 10: Label và Feature của câu lệnh bên dưới có nghĩa là gì?

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 đó được gán vào feature
C. Dữ liệu đầu vào được gán là feature và kết quả của dữ liệu đầu vào đó được gán vào label
D. Dữ liệu đầu vào được gán là label và kết quả dự đoán được gán vào feature