Câu hỏi: AVRO là gì?

A - Avro là một thư viện tuần tự hóa java.

B - Avro là một thư viện nén java.

C - Avro là một thư viện java tạo các tệp bảng chia nhỏ.

D - Không câu trả lời nào đúng.

Câu hỏi: Bản chất của phần cứng cho NameNode phải là

A - Cao cấp hơn loại hàng hóa

B - Loại hàng hóa

C - Không thành vấn đề

D - Chỉ cần có nhiều Ram hơn mỗi DataNode

Câu hỏi: Bản chất DStream:

a, là một chuỗi liên tục RDD

b, Là một chuỗi liên tục DataFrame

c, Là một chuỗi liên tục DataSet

d, ko có đáp án đúng

Câu hỏi: Bạn có thể chạy Map - Reduce jobs trực tiếp trên dữ liệu Avro không?

A - Có, Avro được thiết kế đặc biệt để xử lý dữ liệu qua Map-Reduce.

B - Có, nhưng cần có mã hóa mở rộng bổ sung.

C - Không, Avro được thiết kế đặc biệt chỉ để lưu trữ dữ liệu.

D - Avro chỉ định siêu dữ liệu cho phép truy cập dữ liệu dễ dàng hơn. Dữ liệu này không thể được sử dụng như một phần của quá trình thực thi thu nhỏ bản đồ, thay vì chỉ đặc tả đầu vào.

Câu hỏi: Bạn có thể dự trữ lượng sử dụng đĩa trong một DataNode bằng cách định cấu hình **dfs.datanode.du.reserved** trong tệp nào sau đây

A. Hdfs-site.xml

B. Hdfs-defaukt.xml

C. Core-site.xml

D. Mapred-site.xml

Câu hỏi: Bộ nhớ đệm phân tán là gì?

A - Bộ đệm phân tán là thành phần đặc biệt trên NameNode sẽ lưu vào bộ đệm dữ liệu được sử dụng thường xuyên để phản hồi máy khách nhanh hơn. Nó được sử dụng trong bước giảm.

B - Bộ nhớ đệm phân tán là thành phần đặc biệt trên DataNode sẽ lưu vào bộ đệm dữ liệu được sử dụng thường xuyên để phản hồi máy khách nhanh hơn. Nó được sử dụng trong bước maping.

C - Bộ đệm phân tán là một thành phần lưu trữ các đối tượng java.

D - Bộ nhớ đệm phân tán là một thành phần cho phép các nhà phát triển triển khai các chum để xử lý Map-Reduce.

Câu hỏi: 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 hỏi: Các khối dữ liệu ánh xạ thông tin với các tệp tương ứng của chúng được lưu trữ trong

A - DataNode

B - Job Tracker

C - Task Tracker

D – NameNode

Câu hỏi: 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 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

Câu hỏi: Các tệp HDFS được thiết kế cho

A - Nhiều người viết và sửa đổi ở các hiệu số tùy ý.

B - Chỉ nối vào cuối tệp

C - Chỉ ghi thành tệp một lần.

D - Truy cập dữ liệu có độ trễ thấp.

Câu hỏi: Các ứng dụng người dùng có thể hướng dẫn NameNode để lưu vào bộ đệm các tệp bằng cách

A - thêm tên tệp bộ đệm vào nhóm bộ đệm

B - thêm cấu hình bộ đệm vào nhóm bộ đệm

C - thêm chỉ thị bộ nhớ cache vào nhóm bộ nhớ cache

D - chuyển tên tệp làm tham số cho nhóm bộ nhớ cache

Câu hỏi: Cái nào là một trong những tính năng dữ liệu lớn?

A - Vận tốc Velocity

B - Tính xác thực Veracity

C – khối lượng Volume

D - đa dạng Variety

Câu hỏi: Cái nào sau đây lưu trữ dữ liệu?

A - Name node

B - Data node

C - Master node

D - None of these

Câu hỏi: Cái nào trong số này cung cấp hệ thống xử lý Luồng được sử dụng trong hệ sinh thái Hadoop?

A - Solr

B - Tez

C - Spark

D – Hive

Câu hỏi: 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

Câu hỏi: Chạy Start-dfs.sh kết quả là

A. Bắt đầu NameNode và DataNode

B. Chỉ NameNode bắt đầu

C. Chỉ bắt đầu datanode

D. Khởi động NameNode và trình quản lý tài nguyên

Câu hỏi: Chế độ cài đặt phân phối hoàn toàn (không ảo hóa) cần tối thiểu (The fully distributed mode of installation(without virtualization) needs a minimum of)

A. 2 Máy vật lý

B. 3 Máy vật lý

C. 4 Máy vật lý

D. 1 Máy vật lý

Câu hỏi: Chế độ nào sau đây không phải là chế độ hoạt động của Hadoop?

A - Pseudo distributed mode

B - Globally distributed mode

C - Stand alone mode

D - Fully-Distributed mode

Câu hỏi: Chọn phát biểu đúng khi nói về MongoDB

a, MongoDB có các trình điều khiển driver cho nhiều ngôn ngữ lập trình khác nhau.

b, các văn bản có thể chứa nhiều cặp key-value hoặc key-array, hoặc các văn bản lồng (nested documents)

c, tất cả các phương án trên

d, MongoDB hay các NoSQL có khả năng khả mở tốt hơn các CSDL quan hệ truyền thống

Câu hỏi: Công cụ Hadoop được sử dụng để phân tán dữ liệu một cách **đồng nhất** trên các DataNode được đặt tên là:

A - Scheduler

B - Balancer

C - Spreader

D – Reporter

Câu hỏi: Công nghệ nào được sử dụng để lưu trữ dữ liệu trong Hadoop?

A - HBase

C - Sqoop

B - Avro

D - Zookeeper

Câu hỏi: Công nghệ nào được sử dụng để nhập và xuất dữ liệu trong Hadoop?

A - HBase

C - Sqoop

B - Avro

D – Zookeeper

Câu hỏi: Công nghệ nào được sử dụng để tuần tự hóa dữ liệu trong Hadoop?

A - HBase

C - Sqoop

B - Avro

D – Zookeeper

Câu hỏi: Công nghệ nào sau đây là cơ sở dữ liệu lưu trữ tài liệu?

A - HBase

C - Cassandra

B - Hive

D – CouchDB

Câu hỏi: Công ty nào đã phát triển Apache Cassandra giai đoạn đầu tiên

a, Google

c, linkedin

b, twitter

d, facebook

Câu hỏi: Cơ chế chịu lỗi của datanode trong HDFS

a, dử 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 datanode

Câu hỏi: Cơ chế nào sau đây không phải là cơ chế hàng rào cho NameNode đã hoạt động trước đó?

A - Tắt cổng mạng của nó thơng qua lệnh quản lý từ xa.

B - Thu hồi quyền truy cập của nó vào thư mục lưu trữ được chia sẻ.

C - Định dạng ổ đĩa của nó.

D – STONITH

Câu hỏi: Cơ chế nhân bản dữ liệu trong HDFS

a, Namenode quyết định vị trí các nhân bản của các chunk trên các datanode

b, Datanode là primary quyết định vị trí các nhân bản của cac chunk tại các secondary datanode

c, Client quyết định vị trí lưu trữ các nhân bản với từng chunk

Câu hỏi: 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 data node

c, các chunk được lưu trữ tin cậy trên datanode theo cơ chế RAID

Câu hỏi: DataNode và NameNode là tương ứng

A - Nút chính và nút công nhân

C - Cả hai đều là các nút công nhân

B - Nút công nhân và nút chính

D - Không có

Câu hỏi: Dấu phẩy được sử dụng để sao chép một dạng thư mục từ node này sang node khác trong HDFS là

A. rcp

C. drcp

B. dcp

D. distcp

Câu hỏi: Dữ liệu từ một cụm hadoop từ xa có thể

A. khơng được đọc bởi một cụm hadoop khác

C. được đọc bằng hhtp

B. được đọc bằng http

D. được đọc bằng hftp

Câu hỏi: Đáp án nào không phải là một “output operation ” khi thao tác với DStream

a, saveAsTextFile

c, saveasHadoopFile

b, foreachRDD

d, reduceByKetAndWindow

Câu hỏi: Đáp án nào không phải là một “Transformation” khi thao tác với DStream

a, reduceByWindow

c, foreachWindow

b, window

d, countByWindow

Câu hỏi: Đâu không phải là tính năng mà NoSQL nào cũng đáp ứng

a, tính sẵn sàng cao

c, phù hợp với dữ liệu lớn

b, khả năng mở rộng linh hoạt

Câu hỏi: 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

Câu hỏi: Đâu là lệnh lưu trữ dữ liệu ra ngoài chương trình Spark:

a, input.saveAsTextFile(‘file:///usr/momoinu/mon_loz/hihi.txt’)

b, input.saveAsTextFile(‘/usr/momoinu/mon_loz/hihi.txt’)

c, input.saveAs (‘file:///usr/momoinu/mon_loz/hihi.txt’)

d, input.saveAsTextFile: ‘file:///usr/momoinu/mon_loz/hihi.txt’

Câu hỏi: Đâu là một dạng của NoSQL

a, MySQL

c, Key-value store

b, JSON

d, OLAP.

Câu hỏi: Đầu ra của một map task là

A - Cặp khóa-giá trị của tất cả các bản ghi của tập dữ liệu.

B - Cặp khóa-giá trị của tất cả các bản ghi từ phần tách đầu vào được trình ánh xạ xử lý

C - Chỉ các phím được sắp xếp từ phần tách đầu vào

D - Số hàng được xử lý bởi tác vụ ánh xạ.

Câu hỏi: Để hủy lưu trữ một tệp đã được lưu trữ trong haddop, hãy sử dụng lệnh

A. Unrar

C. Cp

B. Unhar

D. Cphar

Câu hỏi: Điều gì là đúng về HDFS?

A - Hệ thống tệp HDFS có thể được gắn trên Hệ thống tệp của máy khách cục bộ bằng NFS.

B - Hệ thống tệp HDFS không bao giờ có thể được gắn vào Hệ thống tệp của máy khách cục bộ.

C - Bạn có thể chỉnh sửa bản ghi hiện có trong tệp HDFS đã được gắn kết bằng NFS.

D - Bạn không thể thêm vào tệp HDFS được gắn bằng NFS.

Câu hỏi: Điều nào sau đây đúng với ổ đĩa trong một khoảng thời gian?

A - Thời gian tìm kiếm dữ liệu đang cải thiện nhanh hơn tốc độ truyền dữ liệu.

B - Thời gian tìm kiếm dữ liệu đang cải thiện chậm hơn tốc độ truyền dữ liệu.

C - Thời gian tìm kiếm dữ liệu và tốc độ truyền dữ liệu đều đang tăng tương ứng.

D - Chỉ tăng dung lượng lưu trữ mà khong tăng tốc độ truyền dữ liệu.

Câu hỏi: Điều nào sau đây không đúng đối với Hadoop?

A - Đây là một khung phân tán.

B - Thuật toán chính được sử dụng trong đó là Map Reduce

C - Nó chạy với đồ cứng hàng hóa

D - Tất cả đều đúng

Câu hỏi: Điều nào sau đây không phải là mục tiêu của HDFS?

A. Phát hiện lỗi và khôi phục

B. Xử lý tập dữ liệu khổng lồ

C. Ngăn chặn việc xóa dữ liệu

D. Cung cấp băng thơng mạng cao để di chuyển dữ liệu

Câu hỏi: Điều sau không được phép trên các tệp HDFS

A - Xóa

B - Đổi tên

C - Di chuyển

D - Đang thực hiện.

Câu hỏi: Định dạng đầu vào mặc định là gì?

A - Định dạng đầu vào mặc định là xml. Nhà phát triển có thể chỉ định các định dạng đầu vào khác nếu thích hợp nếu xml khơng phải là đầu vào chính xác.

B - Không có định dạng nhập mặc định. Định dạng đầu vào luôn phải được chỉ định.

C - Định dạng đầu vào mặc định là định dạng tệp tuần tự. Dữ liệu cần được xử lý trước trước khi sử dụng định dạng đầu vào mặc định.

D - Định dạng đầu vào mặc định là TextInputFormat với phần bù byte làm khóa và tồn bộ dịng dưới dạng giá trị.

Câu hỏi: Đối với các tệp HDFS được truy cập thường xuyên, các khối được lưu vào bộ nhớ đệm

A - bộ nhớ của DataNode

B - trong bộ nhớ của NameNode

C - Cả A&B

D - Trong bộ nhớ của ứng dụng khách đã yêu cầu quyền truy cập vào các tệp này.

Câu hỏi: Đối với thư mục HDFS, hệ số sao chép (RF) là

A - giống như RF của các tệp trong thư mục đó

B - 0

C-3

D - Không áp dụng.

Câu hỏi: Giao diện org.apache.hadoop.io.Wording khai báo hai phương thức nào? (Chọn 2 câu trả lời.)

public void readFields(DataInput).

public void read(DataInput).

public void writeFields(DataOutput).

public void write(DataOutput).

A - 1 & 4

C - 3 & 4

B - 2 & 3

D - 2 & 4

Câu hỏi: Giao tiếp giữa các quá trình giữa các nút khác nhau trong Hadoop sử dụng (The inter process communication between different nodes in Hadoop uses)

A. REST API **B. RPC** C. RMI D. IP Exchange

Câu hỏi: Giữa Pig và Hive, công cụ nào có giao diện truy vấn gắn với ANSI SQL hơn

A. Pig

C. Hive

B. không phải 2 đáp án trên

Câu hỏi: Hadoop được viết bằng

A - C ++

C - Java

B - Python

D – Go

Câu hỏi: hadoop fs –expunge

A. Cung cấp danh sách các DataNode

B. Được sử dụng để xóa một tệp

C. Được sử dụng để trao đổi một tệp giữa hai DataNode.

D. Dọn sạch thùng rác.

Câu hỏi: Hadoop giải quyết bài toán chịu lỗi thông qua kỹ thuật gì? Chọn đáp án SAI

a, 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 hỏi: Hadoop giải quyết bài toán khả mở bằng cách nào? Chọn đáp án sai

a, Thiết kế 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 vai trò tính toán và lưu trữ

d, Các node thêm vào cụm có thể có cấu hình, độ tin cậy cao

Câu hỏi: Hadoop khác với máy tính tình nguyện ở chỗ

A. Tình nguyện viên đóng góp thời gian CPU chứ không phải băng thông mạng.

B. Tình nguyện viên đóng góp băng thơng mạng chứ không phải thời gian CPU.

C. Hadoop không thể tìm kiếm các số nguyên tố lớn.

D. Chỉ Hadoop mới có thể sử dụng mapreduce.

Câu hỏi: Hadoop sử dụng những cơ chế nào để làm cho namenode có khả năng chống lại sự cố.

A - Sao lưu siêu dữ liệu hệ thống tệp vào đĩa cục bộ và gắn kết NFS từ xa.

B - Lưu trữ siêu dữ liệu hệ thống tệp trên đám mây.

C - Sử dụng máy có ít nhất 12 CPU

D - Sử dụng phần cứng đắt tiền và đáng tin cậy.

Câu hỏi: Hadoop xử lý khối lượng lớn dữ liệu như thế nào?

A - Hadoop sử dụng song song rất nhiều máy. Điều này tối ưu hóa việc xử lý dữ liệu.

B - Hadoop được thiết kế đặc biệt để xử lý lượng lớn dữ liệu bằng cách tận dụng phần cứng MPP.

C - Hadoop gửi mã đến dữ liệu thay vì gửi dữ liệu đến mã.

D - Hadoop sử dụng các kỹ thuật bộ nhớ đệm phức tạp trên NameNode để tăng tốc độ xử lý dữ liệu.

Câu hỏi: HBASE là gì?

A - Hbase là bộ Java API riêng biệt cho cụm Hadoop.

B - Hbase là một phần của dự án Apache Hadoop cung cấp giao diện để quét một lượng lớn dữ liệu bằng cơ sở hạ tầng Hadoop.

C - Hbase là một "cơ sở dữ liệu" giống như giao diện với dữ liệu cụm Hadoop.

D - HBase là một phần của dự án Apache Hadoop cung cấp giao diện giống SQL để xử lý dữ liệu.

Câu hỏi: HDFS có thể được truy cập qua HTTP bằng cách sử dụng

A - lược đồ URI viewfs

C - Lược đồ URI C - wasb

B - lược đồ URI webhdfs

D - HDFS ftp

Câu hỏi: 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

Câu hỏi: HDFS là viết tắt của

A - Hệ thống tệp phân tán cao. (Highly distributed file system.)

B - Hệ thống tệp được hướng dẫn Hadoop (Hadoop directed file system)

C - Vỏ tệp phân tán cao (Highly distributed file shell)

D - Hệ thống tệp phân tán Hadoop. (Hadoop distributed file system.)

Câu hỏi: Hệ số sao chép của tệp trong HDFS có thể được thay đổi bằng cách sử dụng

A. changerep

C. setrep

B. rerep

D. xrep

Câu hỏi: Hệ thống apache nào dưới đây giải quyết việc nhập dữ liệu phát trực tuyến vào hadoop

A - Ozie

C - Flume

B - Kafka

D – Hive

Câu hỏi: Hệ thống nào cho phép đọc ghi dữ liệu tại vị trí ngẫu nhiên, thời gian thực tới hàng terabyte dữ liệu

a, Hbase

c, Pig

b, Flume

d, HDFS

Câu hỏi: Job tracker chạy trên

A - Namenode

C - Secondary namenode

B - Datanode

D - Secondary datanode

Câu hỏi: Khái niệm sử dụng nhiều máy để xử lý dữ liệu được lưu trữ trong hệ thống phân tán không phải là mới.

Máy tính hiệu suất cao (HPC) sử dụng nhiều máy tính để xử lý khối lượng lớn dữ liệu được lưu trữ trong mạng vùng lưu trữ (SAN). So với HPC, Hadoop

A. Có thể xử lý khối lượng dữ liệu lớn hơn.

B. Có thể chạy trên một số lượng máy lớn hơn HPC cluster.

C. Có thể xử lý dữ liệu nhanh hơn với cùng băng thông mạng so với HPC.

D. Không thể chạy các cơng việc tính tốn chun sâu.

Câu hỏi: Khi bạn tăng số lượng tệp được lưu trữ trong HDFS, Bộ nhớ được yêu cầu bởi NameNode

A. Tăng

C. Vẫn khơng thay đổi

B. Giảm

D. Có thể tăng hoặc giảm

Câu hỏi: Khi chạy trên chế độ pseudo distributed (phân phối giả lập), hệ số sao chép được đặt thành

A-2

C-0

B-1

D–3

Câu hỏi: Khi ghi dữ liệu vào HDFS, điều gì là đúng nếu hệ số nhân bản là ba? (Chọn 2 câu trả lời)

1. Dữ liệu được ghi vào DataNodes trên ba giá đỡ riêng biệt (nếu Rack Aware).
2. Dữ liệu được lưu trữ trên mỗi DataNode bằng một tệp riêng biệt chứa checksum.
3. Dữ liệu được ghi vào các khối trên ba DataNodes khác nhau.
4. Khách hàng được trả lại thành công khi ghi thành công khối đầu tiên và kiểm tra tổng kiểm tra.

A - 1 & 3

C - 3 & 4

B - 2 & 3

D - 1 & 4

Câu hỏi: Khi khách hàng giao tiếp với hệ thống tệp HDFS, nó cần giao tiếp với

A - chỉ NameNode

C - cả NameNode và DataNode

B - chỉ DataNode

D - Không có

Câu hỏi: Khi lưu trữ tệp Hadoop, phát biểu nào sau đây là đúng? (Chọn hai câu trả lời)

1. Các tệp đã lưu trữ sẽ hiển thị với phần mở rộng .arc.
2. Nhiều tệp nhỏ sẽ trở thành ít tệp lớn hơn.
3. MapReduce xử lý tên tệp gốc ngay cả sau khi tệp được lưu trữ.
4. Các tệp đã lưu trữ phải được lưu trữ tại Liên hợp quốc cho HDFS và MapReduce để truy cập vào các tệp nhỏ, gốc.
5. Lưu trữ dành cho các tệp cần được lưu nhưng HDFS không còn truy cập được nữa.

A - 1 & 3

C - 2 & 4

B - 2 & 3

D - 3 & 4

Câu hỏi: Khi một jobTracker lên lịch, một công việc sẽ được tìm kiếm đầu tiên

A - Một nút có vị trí trống trong cùng rack với DataNode

B - Bất kỳ nút nào trên cùng rack với DataNode

C - Bất kỳ nút nào trên rack liền kề với rack của datanode

D - Chỉ bất kỳ nút nào trong cụm

Câu hỏi: Khi một máy được khai báo là datanode, dung lượng ổ đĩa trong đó

(When a machine is declared as a datanode, the disk space in it)

A. Chỉ có thể được sử dụng cho lưu trữ HDFS

B. Có thể được sử dụng cho cả lưu trữ HDFS và không phải HDFs

C. Không thể truy cập bằng các lệnh không phải hadoop

D. không thể lưu trữ các tệp văn bản.

Câu hỏi: Khi một node dự phòng được sử dụng trong một cụm thì không cần

A. Node kiểm tra (Check point node)

C. DataNode phụ (Secondary data node)

B. Node tên phụ (Secondary name node)

D. Nhận thức về giá đỡ (Rack awareness)

Câu hỏi: Khi một tệp trong HDFS bị người dùng xóa (When a file in HDFS is deleted by a user)

A. nó đã mất vĩnh viễn

B. Nó sẽ đi vào thùng rác nếu được định cấu hình.

C. Nó bị ẩn khỏi người dùng nhưng vẫn ở trong hệ thống tệp

D. File sin HDFS không thể bị xóa

Câu hỏi: Khi một ứng dụng khách liên hệ với NameNode để truy cập tệp, NameNode phản hồi với

A - Kích thước của tệp được yêu cầu.

B - ID khối của tệp được yêu cầu.

C - ID khối và tên máy chủ của bất kỳ DataNode nào chứa khối đó.

D - Block ID và tên máy chủ của tất cả các DataNode chứa khối đó.

Câu hỏi: Khi NameNode nhận thấy rằng một số khối được sao chép qúa mức, nó

A - Dừng công việc sao chép trong toàn bộ hệ thống tệp hdfs.

B - Nó làm chậm qúa trình nhân bản cho các khối đó

C - Nó xóa các khối thừa.

D - Nó để lại các khối thừa như nó vốn có.

Câu hỏi: Khi sử dụng HDFS, điều gì xảy ra khi tệp bị xóa bởi dòng lệnh?

A - Nó sẽ bị xóa vĩnh viễn nếu thùng rác được bật.

B - Nó được đặt vào một thư mục thùng rác chung cho tất cả người dùng cho cụm đó.

C - Nó bị xóa vĩnh viễn và các thuộc tính tệp được ghi lại trong Editlog.

D - Nó được chuyển vào thư mục thùng rác của người dùng đã xóa nó nếu thùng rác được bật.

Câu hỏi: Kịch bản nào yêu cầu băng thông cao nhất để truyền dữ liệu giữa các nút trong Hadoop?

A - Các nút khác nhau trên cùng một giá đỡ

B - Các nút trên các giá đỡ khác nhau trong cùng một trung tâm dữ liệu.

C - Các nút trong các trung tâm dữ liệu khác nhau

D - Dữ liệu trên cùng một nút.

Câu hỏi: Kích thước khối HDFS lớn hơn so với kích thước của các khối đĩa để

A - Chỉ các tệp HDFS có thể được lưu trữ trong đĩa được sử dụng.

B - Thời gian tìm kiếm là tối đa

C - Không thể chuyển một tệp lớn được tạo từ nhiều khối đĩa.

D - Một tệp duy nhất lớn hơn kích thước đĩa có thể được lưu trữ trên nhiều đĩa trong cụm.

Câu hỏi: 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 được gán vào label

d, dữ liệu đầu vào được gán là lebl và kết quả dự đoán được gán vào feature

Câu hỏi: Là một phần của tính khả dụng cao HDFS, một cặp NameNode chính được cấu hình. Điều gì là đúng với họ?

A - Khi một yêu cầu của khách hàng đến, một trong số họ được chọn ngẫu nhiên sẽ phục vụ yêu cầu đó.

B - Một trong số chúng đang hoạt động trong khi cái còn lại vẫn tắt.

C - Các DataNode chỉ gửi báo cáo khối đến một trong các NameNode.

D - Nút chờ nhận các điểm kiểm tra định kỳ của không gian tên của NameNode đang hoạt động.

Câu hỏi: Lệnh để kiểm tra xem Hadoop có hoạt động hay không là:

A - Jsp

C - Hadoop fs –test

B - Jps

D - Không có

Câu hỏi: Lệnh hadfs được sử dụng để

A - Sao chép tệp từ hệ thống tệp cục bộ sang HDFS.

B - Sao chép tệp hoặc thư mục từ hệ thống tệp cục bộ sang HDFS.

C - Sao chép các tệp từ HDFS sang hệ thống tệp cục bộ.

D - Sao chép tệp hoặc thư mục từ HDFS sang hệ thống tệp cục bộ.

Câu hỏi: Lệnh "hadoop fs -test -z URI" cho kết quả 0 nếu

A. nếu đường dẫn là một thư mục

C. nếu đường dẫn khơng trống

B. nếu đường dẫn là một tệp

D. nếu tệp có độ dài bằng 0

Câu hỏi: Lệnh hdfs để tạo bản sao của tệp từ hệ thống cục bộ là

A - CopyFromLocal

C - CopyLocal

B - copyfromlocal

D – copyFromLocal

Câu hỏi: Lệnh nào liệt kê các khối tạo nên mỗi tệp trong hệ thống tệp.

A - hdfs fsck / -files -blocks

B - hdfs fsck / -blocks -files

C - hdfs fchk / -blocks -files

D - hdfs fchk / -files –block

Câu hỏi: Loại dữ liệu mà Hadoop có thể xử lý là (The type of data Hadoop can deal with is)

A. Structred (Có cấu trúc)

B. Semi-structured (Bán cấu trúc)

C. Unstructured (Khơng có cấu trúc)

D. All of the above (Tất cả những điều trên)

Câu hỏi: Máy khách đọc dữ liệu từ hệ thống tệp HDFS trong Hadoop

A - lấy dữ liệu từ NameNode

B - lấy vị trí khối từ datanode

C - chỉ lấy các vị trí khối tạo thành NameNode

D - lấy cả dữ liệu và vị trí khối từ NameNode

Câu hỏi: Mô tả cach thức một client đọc dữ liệu trên HDFS

a, client thông báo tới namenode để bắt đầu quá trình đọc sau đó client chạy truy vấn các datanode để trực tiếp đọc các chunks

b, client truy vấn Namenodê để biết được vị trí các chunks. Nếu namenode không biết 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

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 vị trí các chunks. Namenode trả về vị trí các chunks. Client kết nối song song tới các datanode để đọc các chunks

Câu hỏi: Một công việc đang chạy trong lon hadoop

A. Bị giết bằng lệnh

B. Không bao giờ có thể bị giết bằng một lệnh

C. Chỉ có thể bị giết bằng cách tắt NameNode

D. Được tạm dừng và chạy lại

Câu hỏi: 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 dữ liệu kiểm tra

d, chạy chương trình dự đoán

Câu hỏi: Mục đích của nút checkpoint trong cụm Hadoop là (The purpose of checkpoint node in a Hadoop cluster is to)

A. Kiểm tra xem NameNode có hoạt động khơng

B. Kiểm tra xem tệp hình ảnh có đồng bộ giữa NodeName và NameNode phụ hay khơng

C. Hợp nhất fsimage và Editlog và tải nó trở lại NameNode đang hoạt động.

D. Kiểm tra xem các DataNode nào không thể truy cập được.

Câu hỏi: Mục đích của sử dụng sparkML là gì

a, chạy MapReduce

b, chạy các thuật toán dự đoán

c, tính toán phân toán

d, cả b và c

Câu hỏi: Mục đích của việc khởi động NameNode trong chế độ khôi phục là để

A. Khôi phục NameNode không thành công

B. Khôi phục một DataNode bị lỗi

C. Khôi phục dữ liệu từ một trong những vị trí lưu trữ siêu dữ liệu

D. Khơi phục dữ liệu khi chỉ có một vị trí lưu trữ siêu dữ liệu

Câu hỏi: Mục tiêu chính của HDFS Tính sẵn sàng cao là

A - Tạo bản sao của NameNode chính nhanh hơn.

B - Để giảm thời gian chu kỳ cần thiết để khôi phục lại NameNode chính mới sau khi nút chính hiện có bị lỗi.

C - Ngăn chặn việc mất dữ liệu do lỗi của NameNode chính.

D - Ngăn chặn biểu mẫu tên chính trở thành điểm lỗi duy nhất.

Câu hỏi: NameNode biết rằng DataNode đang hoạt động bằng cách sử dụng một cơ chế được gọi là

A - heartbeats

C - h-signal

B - datapulse

D - Active-pulse

Câu hỏi: NameNode mất bản sao duy nhất của tệp fsimage. Chúng ta có thể khơi phục điều này từ

A. Datanode

C. Checkpoint node

B. Secondary namenode

D. Never

Câu hỏi: Nếu chúng ta tăng kích thước tệp được lưu trữ trong HDFS mà không tăng số tệp, thì bộ nhớ được yêu cầu bởi NameNode

A. Tăng

C. Vẫn không thay đổi

B. Giảm

D. Có thể tăng hoặc giảm

Câu hỏi: Nếu địa chỉ IP hoặc tên máy chủ của DataNode thay đổi

A - NameNode cập nhật ánh xạ giữa tên tệp và tên khối

B - NameNode không cần cập nhật ánh xạ giữa tên tệp và tên khối

C - Dữ liệu trong DataNode đó sẽ bị mất vĩnh viễn

D - Có NameNode phải được khởi động lại

Câu hỏi: Nguồn của kiến trúc HDFS trong Hadoop có nguồn gốc là (The source of HDFS architecture in Hadoop originated as)

A. Hệ thống tệp phân phối của Google

C. Hệ thống tệp phân tán của Facebook

B. Hệ thống tệp phân tán của Yahoo

D. Hệ thống tệp phân tán Azure

Câu hỏi: Người giữ vườn thú (zookeeper)

A - Phát hiện lỗi của NameNode và chọn NameNode mới.

B - Phát hiện lỗi của các DataNode và chọn một DataNode mới.

C - Ngăn phần cứng quá nóng bằng cách tắt chúng.

D - Duy trì danh sách tất cả các thành phần địa chỉ IP của cụm Hadoop.

Câu hỏi: Nhận thức về rack trong NameNode có nghĩa là (Rack awareness in name node means)

A. Nó biết có bao nhiêu rack có sẵn trong cụm

B. Nó nhận thức được ánh xạ giữa nút và giá đỡ

C. Nó nhận biết được số lượng nút trong mỗi rack

D. Nó biết những DataNode nào khong có sẵn trong cụm.

Câu hỏi: Nhiệm vụ nào sau đây là trong số các nhiệm vụ của các Datanode trong HDFS?

A - Duy trì cây hệ thống tệp và siêu dữ liệu cho tất cả các tệp và thư mục.

B - Không có phương án nào đúng.

C - Kiểm soát việc thực hiện một tác vụ bản đồ riêng lẻ hoặc một tác vụ thu gọn.

D - Lưu trữ và truy xuất các khối khi được khách hàng hoặc NameNode yêu cầu.

E - Quản lý không gian tên hệ thống tệp.

Câu hỏi: Nút nào sau đây quản lý các nút khác?

A - Name node

C - slave node

B - Data node

D - None of these

Câu hỏi: 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 cho phép xử lý kết tập dữ liệu mà kích thước lớn hơn kích thước bộ nhớ trong

c, Presto có cơ chế chịu lỗi khi thực thi truy vấn

Câu hỏi: 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 hỏi: phát biểu nào không đúng vè Apache Hadoop

a, xử lý dữ liệu phân tán với mô hình lập trình đơn giản, than thiện hơn như MapReduce

b, Hadoop thiết kế để mở rộng thông qua kỹ thuật scale-outm tăng số lượng máy chủ

c, thiết kế để vận hàn trên phần cứng phổ thống, có khả năng chống chịu lỗi phần cứng

d, thiết kế đẻ vận hành trên siêu máy tính, cấu hình mạnh, độ tin cậy cao

Câu hỏi: 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 thích hợp với các công cụ Business Intelligence

c, presto được quản lý bởi presto software foundation

d, presto được quản lý bởi apache software foundation

Câu hỏi: Phát biểu nào sau đây sai về Kafka

a, nhiều consumer có thể cùng đọc 1 topic

b, 1 message có thể được đọc bởi nhiều consumer khác nhau

c, số lượng consumer phải ít hơn hoặc bằng số lượng partitions

d, 1 message chỉ có thể được đọc bởi 1 consumer trong 1 consumer group

Câu hỏi: phát biểu nào sau đây sai về kafka

a, partition được nhân bản ra nhiều brokers

b, message sau khi được tiêu thụ (consume) thì không bị xóa.

c, các topic gồm nhiều partitions

d, Kafka đảm bảo thứ tự của các message với mỗi topics

Câu hỏi: Sao chép thiếu trong HDFS có nghĩa là

A - Không có sự sao chép nào diễn ra trong các DataNode.

B - Quá trình sao chép rất chậm trong các DataNode.

C - Tần suất sao chép trong các DataNode rất thấp.

D - Số lượng bản sao được tái tạo ít hơn so với quy định của hệ số sao chép.

Câu hỏi: So với RDBMS, Hadoop

A. Có tính tồn vẹn dữ liệu cao hơn.

B. Có giao dịch ACID không

C. IS thích hợp để đọc và viết nhiều lần

D. Hoạt động tốt hơn trên dữ liệu phi cấu trúc và bán cấu trúc.

Câu hỏi: Số lượng nhiệm vụ mà trình theo dõi tác vụ có thể chấp nhận phụ thuộc vào

A. Bộ nhớ tối đa có sẵn trong nút

B. Không giới hạn

C. Số lượng slot được định cấu hình trong đó

D. Theo quyết định của jobTracker

Câu hỏi: 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

Câu hỏi: Spark hỗ trợ các cluster manager nào

a, Standalone Cluster manager

b, MESOS

c, YARN

d, tất cả đáp án trên

Câu hỏi: Spark Streaming trừu tượng hóa cũng như thao tác với các dòng dữ liệu (data stream) dựa trên khái niệm nào:

a, shared variable

b, RDD

c, DStream

d, DataFrame

Câu hỏi: Sự khác biệt giữa chế độ standalone and pseudo-distributed mode là

A - Standalone không thể sử dụng map reduce

B - Standalone có một quy trình java duy nhất chạy trong đó.

C - pseudo-distributed mode không sử dụng HDFS

D - pseudo-distributed mode cần hai hoặc nhiều máy vật lý.

Câu hỏi: Sự phân chia đầu vào được sử dụng trong MapReduce cho biết

A - Kích thước trung bình của các khối dữ liệu được sử dụng làm đầu vào cho chương trình

B - Chi tiết vị trí nơi bắt đầu của toàn bộ bản ghi đầu tiên trong một khối và toàn bộ bản ghi cuối cùng trong khối kết thúc.

C - Tách dữ liệu đầu vào cho chương trình MapReduce thành kích thước đã được định cấu hình trong mapred-site.xml

D - Không có

Câu hỏi: Tất cả các tệp trong một thư mục trong HDFS có thể được hợp nhất với nhau bằng cách sử dụng

A. Getmerge

C. remerge

B. putmerge

D. mergeall

Câu hỏi: Tệp HDFS nhỏ hơn kích thước một khối

A - Không thể lưu trữ trong HDFS.

B - Chiếm tồn bộ kích thước của khối.

C - Chỉ chiếm kích thước mà nó cần chứ không phải tòan khối.

D - Có thể trải dài trên nhiều khối.

Câu hỏi: Tệp lưu trữ được tạo trong Hadoop ln có phần mở rộng là

A. hrc

C. Hrh

B. Har

D. Hrar

Câu hỏi: Tệp trong Namenode lưu trữ thông tin ánh xạ vị trí khối dữ liệu với tên tệp là

A - dfsimage

C - fsimage

B - nameimage

D – image

Câu hỏi: Thành phần nào không thuộc thành phần lõi của Hadoop

a, Hệ thống tập tin phân tán HDFS

d, Apache ZooKeeper

b, MapReduce Framework

e, Apache Hbase

c, YARN: yet another resource negotiator

Câu hỏi: Thành phần nào sau đây truy xuất các phần đầu vào trực tiếp từ HDFS để xác định số tác vụ map?

A - Namenode.

D - JobTracker.

B - TaskTrackers.

E - Không có lựa chọn nào đúng.

C - JobClient.

Câu hỏi: Theo Tính khả dụng cao của Hadoop, nghĩa là Hàng rào

A - Ngăn NameNode hoạt động trước đó bắt đầu chạy lại.

B - Ngăn chặn việc bắt đầu chuyển đổi dự phòng trong trường hợp mạng bị lỗi với NameNode hoạt động.

C - Ngăn chặn sự cố sập nguồn đối với NameNode đã hoạt động trước đó.

D - Ngăn không cho NameNode đã hoạt động trước đó ghi và Editlog.

Câu hỏi: Thuộc tính được sử dụng để đặt hệ thống tệp mặc định cho Hadoop trong

core-site.xml là

A - filesystem.default

B - fs.default

C - fs.defaultFS

D - hdfs.default

Câu hỏi: Thuộc tính nào dưới đây được định cấu hình trên core-site.xml?

A – hệ số nhân bản

B - Tên thư mục để lưu trữ tệp hdfs.

C - Máy chủ và cổng nơi tác vụ MapReduce chạy.

D - Các biến môi trường Java.

Câu hỏi: Thuộc tính nào dưới đây được định cấu hình trên hadoop-env.sh?

A – Hệ số nhân bản

B - Tên thư mục để lưu trữ tệp hdfs

C - Máy chủ và cổng nơi tác vụ MapReduce chạy

D - Các biến môi trường Java.

Câu hỏi: Thuộc tính nào dưới đây được định cấu hình trên hdfs-site.xml?

A – Hệ số nhân bản

B - Tên thư mục để lưu trữ tệp hdfs.

C - Máy chủ và cổng nơi tác vụ MapReduce chạy.

D - Các biến môi trường Java.

Câu hỏi: Thuộc tính nào dưới đây được định cấu hình trên mapred-site.xml?

A – Hệ số nhân bản

B - Tên thư mục để lưu trữ tệp hdfs.

C - Máy chủ và cổng nơi tác vụ MapReduce chạy.

D - Các biến môi trường Java.

Câu hỏi: Tiện ích nào được sử dụng để kiểm tra tình trạng của hệ thống tệp HDFS?

A - fchk **B - fsck** C - fsch D – fcks

Câu hỏi: Tín hiệu heartbeat được gửi từ

A - JObtracker thành Tasktracker

B - Tasktracker tới Job tracker

C - Trình theo dõi công việc đến NameNode

D - Trình theo dõi tác vụ đến NameNode

Câu hỏi: Tính năng decommission trong hadoop được sử dụng cho

A. Hủy cấp phép NameNode C. Hủy cấp phép NameNode phụ.

B. Hủy khai thác các DataNode D. Giải nén tồn bộ cụm Hadoop

.

Câu hỏi: Tính năng định vị dữ liệu trong Hadoop có nghĩa là

A - lưu trữ cùng một dữ liệu trên nhiều nút.

B - chuyển vị trí dữ liệu từ nút này sang nút khác.

C - đồng định vị dữ liệu với các nút tính tóan.

D - Phân phối dữ liệu trên nhiều nút.

Câu hỏi: Trong đĩa cục bộ của NameNode, các tệp được lưu trữ liên tục là:

A - fsimage và editlog

B - vị trí khối và hình ảnh vùng tên

C - chỉnh sửa nhật ký và chặn vị trí

D - Hình ảnh không gian tên, chỉnh sửa vị trí nhật ký và chặn.

Câu hỏi: Trong Hadoop 2.x, liên kết HDFS phát hành có nghĩa là

A - Cho phép các NameNode giao tiếp với nhau.

B - Cho phép một cụm mở rộng quy mô bằng cách thêm nhiều DataNode dưới một NameNode.

C - Cho phép một cụm mở rộng quy mô bằng cách thêm nhiều NameNode hơn.

D - Thêm nhiều bộ nhớ vật lý hơn cho cả NameNode và DataNode.

54. Theo liên kết HDFS

A - Mỗi NameNode quản lý siêu dữ liệu của toàn bộ hệ thống tệp.

B - Mỗi NameNode quản lý siêu dữ liệu của một phần hệ thống tệp.

C - Lỗi một NameNode làm mất một số tính khả dụng của siêu dữ liệu từ toàn bộ hệ thống tệp.

D - Mỗi DataNode đăng ký với mỗi NameNode.

Câu hỏi: Trong Hadoop, Snappy và LZO là những ví dụ về

A - Cơ chế vận chuyển tệp giữa các DataNode

B - Cơ chế nén dữ liệu

C - Cơ chế sao chép dữ liệu

D - Cơ chế đồng bộ hóa dữ liệu

Câu hỏi: Trong HDFS, các tệp không thể

A. Đọc

B. Xóa

C. Thực thi

D. Lưu trữ (Archived)

Câu hỏi: Trong hệ sinh thái của Spark không có công cụ hay thành phần nào sau đây:

a, MLib

b, GraphX

c, Sqoop

d, Cluster Managers

Câu hỏi: Trong hệ thống HDFS với kích thước khối 64MB, chúng tôi lưu trữ một tệp nhỏ hơn 64MB. Điều nào sau đây là đúng?

A. Tệp sẽ tiêu tốn 64MB

C. Tệp sẽ tiêu tốn ít hơn 64MB.

B. Tệp sẽ tiêu tốn hơn 64MB

D. Không thể đoán trước được.

Câu hỏi: Trong một cụm Hadoop, điều gì đúng với khối HDFS không còn khả dụng do hỏng đĩa hoặc lỗi máy?

A - Nó bị mất vĩnh viễn

B - Nó có thể được sao chép ở các vị trí thay thế của nó cho các máy sống khác.

C - NameNode cho phép yêu cầu của khách hàng mới tiếp tục cố gắng đọc nó.

D - Tiến trình công việc Mapreduce chạy bỏ qua khối và dữ liệu được lưu trữ trong đó.

Câu hỏi: Trong Secondary NameNode, lượng bộ nhớ cần thiết là

A. Tương tự như của node chính

B. Phải có ít nhất một nửa node chính

C. Phải gấp đơi node chính

D. Chỉ phụ thuộc vào số lượng node dữ liệu mà nó sẽ xử lý

Câu hỏi: Tùy chọn nào sau đây không phải là tùy chọn lập lịch có sẵn trong YARN

A - Balanced scheduler (lập lịch cân bằng)

C - Capacity scheduler (Lập lịch dung lượng)

B - Fair scheduler (Lập lịch công bằng)

D - FiFO schesduler.

Câu hỏi: Tùy chọn nào trong số này không phải là tùy chọn lập lịch có sẵn với YARN?

A - Optimal Scheduler (lập lịch tối ưu)

C - Capacity scheduler

B - FIFO scheduler

D - Fair scheduler

Câu hỏi: Vai trị chính của NameNode phụ là

A - Sao chép siêu dữ liệu hệ thống tệp từ NameNode chính.

B - Sao chép siêu dữ liệu hệ thống tệp từ NFS được lưu trữ bởi NameNode

chính

C - Theo dõi xem NameNode chính có đang hoạt động hay khơng.

D - Định kỳ hợp nhất fsimage với editlog.

Câu hỏi: Vai trò của Journal node là

A - Báo cáo vị trí của các khối trong một DataNode

B - Báo cáo thông tin editlog của các khối trong DataNode.

C - Báo cáo lịch trình khi công việc sẽ chạy

D - Báo cáo hoạt động của các thành phần khác nhau do người quản lý tài nguyên xử lý

Câu hỏi: Vấn đề chính gặp phải khi đọc và ghi dữ liệu song song từ nhiều đĩa là gì?

A - Xử lý khối lượng lớn dữ liệu nhanh hơn.

B - Kết hợp dữ liệu từ nhiều đĩa.

C - Phần mềm cần thiết để thực hiện nhiệm vụ này là cực kỳ tốn kém.

D - Phần cứng cần thiết để thực hiện tác vụ này là cực kỳ tốn kém.

Câu hỏi: Vị trí khối hiện tại của HDFS nơi dữ liệu đang được ghi vào,

A - hiển thị cho khách hàng yêu cầu nó.

B - Vị trí khối không bao giờ hiển thị đối với các yêu cầu của khách hàng.

C - Người đọc có thể nhìn thấy hoặc không.

D - chỉ hiển thị sau khi dữ liệu được lưu trong bộ đệm được cam kết.

Câu hỏi: Writable là gì?

A - W ghi là một giao diện java cần được triển khai để truyền dữ liệu trực tuyến đến các máy chủ từ xa.

B - W ghi là một giao diện java cần được thực hiện để ghi HDFS.

C - Writes là một giao diện java cần được triển khai để xử lý MapReduce

D - Không câu trả lời nào đúng.

Câu hỏi: YARN là viết tắt của

A. Yahoo’s another resource name

C. Yahoo’s archived Resource names

B. Yet another resource negotiator

D. Yet another resource need.

Câu hỏi: Yếu tố giới hạn hiện tại đối với kích thước của một cụm hadoop là

A. Nhiệt lượng dư thừa tạo ra trong trung tâm dữ liệu

B. Giới hạn trên của băng thông mạng

C. Giới hạn trên của RAM trong NameNode

D. 4000 datanode

Câu hỏi: Yếu tố sao chép mặc định cho hệ thống tệp HDFS trong hadoop là

A-1 B-2 **C-3** D-4

Câu hỏi: Zookeeper đảm bảo rằng

A - Tất cả các NameNode đang tích cực phục vụ các yêu cầu của khách hàng

B - Chỉ có một NameNode đang tích cực phục vụ các yêu cầu của khách hàng

C - Chuyển đổi dự phịng được kích hoạt khi bất kỳ DataNode nào bị lỗi.

D - Quản trị viên hadoop khơng thể bắt đầu chuyển đổi dự phòng.

122. Câu nào sau đây là đúng đối với các cặp <key, value> của một công việc MapReduce?

A - Một lớp khóa phải triển khai Words.

B - Một lớp khóa phải triển khai WordsComp so sánh được.

C - Một lớp giá trị phải triển khai WordsComp so sánh được.

D - Một lớp giá trị phải mở rộng khả năng so sánh được.

123. Phát biểu nào sau đây là sai về Bộ nhớ đệm phân tán?

A - Khung công tác Hadoop sẽ đảm bảo rằng bất kỳ tệp nào trong Bộ đệm phân tán được phân phối cho tất cả các tác vụ bản đồ và giảm bớt.

B - Các tệp trong bộ đệm có thể là tệp văn bản hoặc chúng có thể là tệp lưu trữ như tệp zip và JAR.

C - Disk I / O bị tránh vì dữ liệu trong bộ đệm được lưu trong bộ nhớ.

D - Khung công tác Hadoop sẽ sao chép các tệp trong Bộ đệm ẩn phân tán vào nút phụ trước khi bất kỳ nhiệm vụ nào cho công việc được thực thi trên nút đó.

124. Thành phần nào sau đây **không** phải là thành phần chính của HBase?

A - Máy chủ Vùng.

C - ZooKeeper.

B - Nagios.

D - Máy chủ chính.

125. Điều nào sau đây là sai về RawComparator?

A - So sánh các khóa theo byte.

B - Hiệu suất có thể được cải thiện trong giai đoạn sắp xếp và đủ bằng cách sử dụng RawComparator.

C - Các khóa trung gian được giải mã hóa để thực hiện so sánh.

126. Con quỷ (demon) nào chịu trách nhiệm sao chép dữ liệu trong Hadoop?

A - HDFS.

D - NameNode.

B - Trình theo dõi tác vụ.

E - DataNode.

C - Trình theo dõi công việc.

127. Các phím từ đầu ra của xáo trộn và sắp xếp thực hiện giao diện nào sau đây?

A - Viết được.

D - Có thể so sánh được.

B - Có thể so sánh được.

E - Có thể so sánh được.

C - Có thể cấu hình.

128. Để áp dụng một bộ kết hợp, một thuộc tính phải được thỏa mãn bởi các giá trị được phát ra từ bộ ánh xạ là gì?

A - Combiner luôn có thể được áp dụng cho mọi dữ liệu

B - Đầu ra của bộ ánh xạ và đầu ra của bộ kết hợp phải cùng một cặp giá trị khóa và chúng có thể không đồng nhất

C - Đầu ra của bộ ánh xạ và đầu ra của bộ kết hợp phải cùng một cặp giá trị khóa. Chỉ khi các giá trị thỏa mãn thuộc tính liên kết và giao hoán thì nó mới có thể được thực hiện

1. Phát biểu về định lý CAP

The limitations of distributed databases can be described in the so called the CAP theorem

• Consistency: every node always sees the same data at any given instance (i.e., strict consistency)

• Availability: the system continues to operate, even if nodes in a cluster crash, or some hardware or software parts are down due to upgrades

• Partition Tolerance: the system continues to operate in the presence of network partitions

2. Giải thích ngắn gọn về các phương pháp tổng quát trong việc thích nghi các giải thuật học máy cho dữ liệu lớn

3. Cho đoạn chương trình sau:

rdd = sc.parallelize([“hello”, “world”, “good”, “hello”], 2)

rdd = rdd.map(lambda w : (w, 1))

Đưa ra kết quả của rdd.glom().collect()

import numpy as np

import os

packages = "org.apache.spark:spark-sql_2.12:3.0.1"

os.environ["PYSPARK_SUBMIT_ARGS"] = (

```
"--packages {0} pyspark-shell".format(packages)
)
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
from pyspark.sql import SQLContext
sc = SparkContext('local')
spark = SparkSession(sc)
sqlContext = SQLContext(sc)
rdd = sc.parallelize(["hello", "world", "good", "hello"], 2)
rdd = rdd.map(lambda w : (w, 1))
result=rdd.glom().collect()
print(result)
```

4. Giả sử một textfile kích thước lớn được đặt ở đường dẫn hdfs://user/bigfile.txt. Viết chương trình Spark đếm số lần xuất hiện của chuỗi “big data processing” trong file text này.

```
linesRDD = sc.textFile("hdfs://user/bigfile.txt")
df=spark.read.format("json").load("../data/flight-data/json/2015-summary.json")
df = (spark.read.format("com.databricks.spark.csv")
.option("header", "true")
#.option("inferSchema","true")
.schema(schema)
.load("hdfs://192.168.56.10:9000/user/output/data10/*.csv"))
df.printSchema()
df.createOrReplaceTempView("dfTable")
print("Số bản ghi:")
print(df.count())
df.show(20, False)
```

5. Cho đoạn chương trình sau

```
def mystr(d):
return d
def myconcat(a, b):
return a + b
def mypartConcat(a, b):
return a + b
rdd = sc.parallelize([ (“a”, 1), (“b”, 1), (“a”, 2), (“a”, 8), (“c”, 4), (“a”, 12), (“a”, 18), (“c”, 14) ],
3)
```

Tinhs kết quả của rdd.combineByKey(mystr, myconcat,

(đcm câu này thiếu dòng cuối, gõ đến đây ms thấy)