



Big Data with note

Các câu Presto, Hbase, Kafka, không học: 5, 6, 7, 8, 12, 13, 14, 15, 18, 24, 33, 35, 36, 37

Question 1: 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. Không thể sử dụng SQL để truy vấn dữ liệu NoSQL

Có thể sử dụng ngôn ngữ SQL-like để truy vấn

☒ c. Mở rộng theo chiều ngang, tinh chỉnh được tính sẵn sàng của hệ thống

NoSQL mở rộng ngang (Scale out - Tăng phần cứng) tốt hơn RDBMS

☐ d. Mở rộng theo chiều dọc, thiết kế đơn giản, khó tinh chỉnh tính sẵn sàng của hệ thống

RDBMS mới mở dọc (Scale up)



Ngoài ra 1 số đặc điểm khác của NoSQL

- Dùng tính toán phân tán
- Schema less
- Có thể xử lý có un/semi structure
- Các bảng không có quan hệ với nhau như RDBMS
- Khả năng scale out tốt
- Hợp với dl lớn

Question 2: Cơ chế nhân bản dữ liệu trong HDFS?

- ☐ a. Client quyết định vị trí lưu trữ các nhân bản với từng chunk.
- ☐ b. Datanode là primary quyết định vị trí các nhân bản của các chunk tại các secondary datanode.
- ☒ c. Namenode quyết định vị trí các nhân bản của các chunk trên datanode.



Về nhân bản dữ liệu trong HDFS

- Kích thước Block và số lượng nhân bản có thể tùy chỉnh cho từng file
- Số lượng nhân bản được xác định tại thời điểm tạo file, có thể thay đổi sau đó
- Namenode quyết định mọi thứ liên quan đến việc nhân bản (Vậy nên A và B sai đó babe)

Question 3: Các mục tiêu chính của Apache Hadoop?

- ☐ a. Xử lý dữ liệu lớn mạnh mẽ
- ☐ b. Trực quan hoá dữ liệu hiệu quả
- ☐ c. Lưu trữ dữ liệu khả mở
- ☒ 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

Question 4: Thành phần nào không thuộc thành phần lõi của Hadoop?

- ☐ a. Mapreduce framework
- ☒ b. Apache Zookeeper
- ☐ c. Hệ thống tệp tin phân tán HDFS
- ☐ d. YARN: yet another resource negotiator
- ☒ e. Apache Hbase



Các thành phần core Hadoop

- HDFS
- MapReduce
- YARN
- Hadoop Common (Cần confirm lại)

Question 5: 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.

Question 6: Phát biểu nào đúng về Amazon DynamoDB

- ☒ a. DynamoDB là zero-hop DHT
- ☐ b. DynamoDB là one-hop DHT
- ☐ c. DynamoDB là multi-hop DHT

Question 7: Phát biểu nào sai về Presto

- ☐ a. Presto có thể truy vấn nhiều data storages khác nhau như HDFS, Cassandra
- ☐ b. Presto thường nhanh hơn Hive hay Pig
- ☒ c. Presto không truy vấn được dữ liệu trong MySQL, MS SQL và các CSDL quan hệ truyền thống

Question 8: Thao tác nào không được hỗ trợ bởi Hbase

- ☐ a.Scan
- ☐ b.Multiput
- ☐ c.Put
- ☒ d.Join
- ☐ e.Get



Các thao tác được hỗ trợ bởi HBase

- Get
- Put
- Delete
- Scan
- Increment

Question 9: Ưu điểm của 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 hoá việc chia sẻ dữ liệu.
- ☒ c.Tính khả mở cao.

Question 10: Đây là một dạng của NoSQL

- ☐ a.MySQL
- ☒ b.JSON
- ☒ c.Key-value-store
- ☐ d.OLAP



Các dạng NoSQL

- Key-value pair (Key-value store): Memcached, Redis, DynamoDB
- Column-oriented (Column-family): BigTable, Hbase, Cassandra
- Graph-based: Neo4j, dex, orientedDB, RDF, Sones, tất cả mọi thứ có chữ “Graph” trong đó
- Document-oriented (Document store): MongoDB, CouchDB, JSON

Question 11: Đây là vấn đề khi xử lý dữ liệu lớn với MapReduce?

- ☒ a. ~~Xử lý dữ liệu lớn trong thời gian tương tác~~
- ☐ b. Xử lý luồng dữ liệu lớn
- ☒ c. ~~Xử lý chuỗi các công việc~~
- ☐ d. Xử lý dữ liệu lớn theo lô (Bulk processing)



Các cons của MapReduce


- Xử lý thời gian thực
- Xử lý dữ liệu streaming (MR hợp batch)
- Tính toán phục vụ OLTP (MR hợp OLAP)

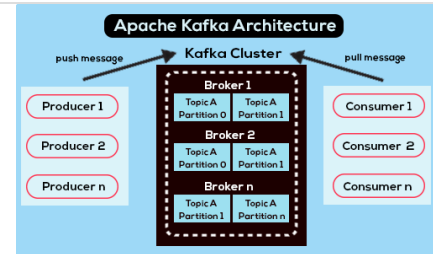
Question 12: Phát biểu nào sai về Kafka?

- ☐ a. Kafka producer có thể gửi message đến nhiều broker khác nhau.
- ☒ b. ~~Thứ tự của message trong mỗi partition do key của message quyết định.~~
(Debate)
- ☒ c. ~~Kafka producer quyết định message sẽ được gửi đến partition nào trong topic.~~
(Debate)

Apache Kafka Architecture and Its Components -The A-Z Guide

Apache Kafka Architecture with Diagram - Explore the Event Driven Architecture of Kafka Cluster, its features and the role of various components.

 <https://www.projectpro.io/article/apache-kafka-architecture-/442>



Question 13: Kiến trúc xử lý dữ liệu Lambda có đặc điểm gì?

- ☐ a. 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.
- ☐ b. Bao gồm các tiến trình ETL (extract, transform, load) đưa dữ liệu vào hồ dữ liệu (data lake)
- ☐ c. Có kiến trúc gồm 2 tầng: tầng xử lý theo lô và tầng xử lý theo luồng
- ☒ d. Kết hợp xử lý dữ liệu theo lô và theo luồng
- ☒ e. 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

Question 14: Phát biểu nào sau sai về Kafka?

- ☒ a. Tiến trình quảng bá message lên cụm Kafka gọi là publishers. (Debate)
- ☐ b. Các máy chủ chạy Kafka gọi là các brokers.
- ☐ c. Tiến trình đăng ký theo dõi các topics gọi là consumers. (Debate)
- ☐ d. Kafka quản lý các luồng thông điệp (messages) thành các nhóm gọi là các Topics.



Thành phần của Kafka

- Topic: Luồng dữ liệu thuộc 1 category/tên feed nhất định
- Brokers: Các server trong cụm kafka
- Consumers: Đọc (Đăng ký) dữ liệu từ cụm
 - Consumer group là các consumer đọc từ 1 topic
- Producers: Viết tin đến 1/nhiều topic
- Partition: Chia nhỏ dữ liệu trong topic thành nhiều phần để đọc viết song song
- Partition offset: Dùng để xác định vị trí 1 record trong partition
- Replicas: Nhân bản backup cho partition
- Leader: 1 server thực hiện toàn bộ tác vụ đọc/viết tại 1 partition
- Follower: Sao chép dữ liệu của leader để lên làm thay khi leader chết

Question 15: Phát biểu nào đúng về Presto?

- ☐ a. Presto có cơ chế chịu lỗi khi thực thi truy vấn
- ☒ b. Các stage được thực thi theo cơ chế pipeline, không có thời gian chờ giữa các stage như Map Reduce
- ☐ c. Presto cho phép xử lý kết tập dữ liệu mà kích thước lớn hơn kích thước bộ nhớ trong

Question 16: Phát biểu nào sau đây không đúng về Apache Hadoop?

- ☐ a. Hadoop thiết kế để mở rộng thông qua kỹ thuật scale-out, tăng số lượng máy chủ
- ☒ b. Thiết kế để vận hành trên siêu máy tính, cấu hình mạnh, độ tin cậy cao
- ☐ c. Xử lý dữ liệu phân tán với mô hình lập trình đơn giản, thân thiện hơn như MapReduce.

- ☐ d. Thiết kế để vận hành trên phần cứng phổ thông, có khả năng chống chịu lỗi phần cứng

Question 17: Hadoop giải quyết bài toán chịu lỗi thông qua kỹ thuật gì. Chọn đáp án sai.

- ☐ a. Các công việc cần tính toán được phân mảnh thành các tác vụ độc lập.
- ☐ b. Hadoop chịu lỗi thông qua kỹ thuật dư thừa
- ☒ c. Các tệp tin được phân mảnh, các mảnh được lưu trữ tin cậy trên ổ cứng theo cơ chế RAID
- ☐ d. Các tệp tin được phân mảnh, các mảnh được nhân bản ra các node khác trên cụm

Question 18: Phát biểu nào sai về Hfile trong Hbase?

- ☐ a. Nhiều Hfile có thể được gộp lại thành 1 Hfile lớn theo những khoảng thời gian nhất định
- ☒ b. Một version của 1 dòng hay 1 bản ghi trong Hbase table có thể được phân rã trên nhiều Hfile khác nhau
- ☐ c. Hfile chứa một tập hợp các dòng bản ghi trong Hbase table
- ☐ d. Nhiều Hfile có thể được gộp lại thành 1 Hfile lớn khi cần thiết

Question 19: 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



Về Spark Transformation

- Input: RDD
- Output: 1 hay nhiều RDD
- Không thay đổi input vì RDD immutable
- Bản chất là lazy: Chỉ biến đổi khi gọi 1 action
 - Action tạo ra RDD nhỏ hơn: filter, count, distinct, sample
 - Action tạo ra RDD to hơn: flatmap, union, cartesian
 - Action tạo ra RDD same size: map
- 2 loại transformation
 - Hẹp: Toàn bộ dữ liệu cần để tính toán tại 1 partition nằm trong 1 partition: map(), filter()
 - Rộng: Cần dl từ nhiều partition để tính toán tại 1 partition

Question 20: Đây 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~~



Spark hơn

- Spark nhanh hơn 100 lần, có thể xử lý CẬN thời gian thực
- Spark có thể sử dụng RAM (bộ nhớ trong)
- Spark có thể hoạt động với dl nhỏ (Không nói là tốt)
- Spark có nhiều API → Dễ lập trình hơn
- Hỗ trợ chuỗi biến đổi



Hadoop hơn

- Hadoop chịu lỗi tốt + an toàn hơn Spark (Oh wow cái này mới)
- Hadoop chi phí rẻ hơn (Disk thay vì RAM)
- Hadoop hợp với dl lớn (Batch)

Question 21: Đâu không phải là tính năng mà NoSQL nào cũng đáp ứng

- ☐ a. Khả năng mở rộng linh hoạt
- ☐ b. Phù hợp với dữ liệu lớn
- ☒ c. Tính sẵn sàng cao



CAP theorem

- CA: MySQL, mấy cái graph
- CP: MongoDB, HBase, Redis
- AP: CouchDB, Cassandra, Riak

Question 22: Các đặc trưng 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 **(Debate)**
- ☐ c. Hỗ trợ nén dữ liệu để tiết kiệm chi phí **(Debate)**
- ☐ d. Tối ưu cho các tệp tin có kích thước lớn



Các đặc trưng HDFS

- Có phân quyền người dùng
- Có cây phân cấp (Hierarchy tree)
- Yêu cầu phần cứng thấp
- Tối ưu tệp lớn
- Viết 1 lần đọc nhiều lần
- Data locality
- Có nén file



Về đọc viết trên HDFS

- Chỉ có 1 client có thể append vào 1 file tại 1 thời điểm
- 1 Datanode có khả năng đọc và viết cùng lúc

Question 23: Giữa Pig và Hive, công cụ nào có giao diện truy vấn gần với ANSI SQL hơn?

- ☐ a.Pig
- ☐ b.Pig và Hive đều không có giao diện truy vấn gần với SQL.
- ☒ c.Hive

Question 24: 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.Split được gán cho task theo cơ chế lazy
- ☐ d.Stage có thể được lập lịch theo giai đoạn

Question 25: 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. Tập trung hoá việc quản trị. **(Debate)**
- ☒ c. Đơn giản hoá việc chia sẻ dữ liệu.

Question 26: HDFS giải quyết bài toán một điểm hỏng hóc duy nhất (single-point-of-failure) cho Namenode bằng cách nào?

- ☒ a. Sử dụng Secondary namenode theo cơ chế active-passive. Secondary namenode chỉ hoạt động khi có vấn đề với Namenode.
- ☐ b. Sử dụng thêm secondary namenode theo cơ chế active-active. Cả Namenode và Secondary namenode cùng online trong hệ thống



Cơ chế hoạt động của secondary namenode

- Name node xin sẽ viết metadata vào bộ nhớ
- Namenode xin chết → 2nd namenode chạy thay, nhưng không có quyền viết vào bộ nhớ nói trên, mà lưu ở 1 chỗ riêng
- Namenode xin quay lại, copy các metadata từ 2nd namenode vào bộ nhớ và chạy tiếp

Question 27: 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 hay các NoSQL có khả năng khả mở tốt hơn các CSDL quan hệ truyền thống.
- ☐ c. MongoDB có các trình điều khiển driver cho nhiều ngôn ngữ lập trình khác nhau.
- ☒ d. Tất cả các phương án đã đưa ra.

Question 28: Cơ chế chịu lỗi của datanode trong HDFS?

- ☒ a. Sử dụng cơ chế heartbeat, định kỳ các datanode thông báo về trạng thái cho Namenode.

- ☐ b.Sử dụng Zookeeper để quản lý các thành viên datanode trong cụm.
- ☐ c.Sử dụng cơ chế heartbeat, Namenode định kỳ hỏi các datanode về trạng thái tồn tại của các datanode.

Question 29: Phát biểu nào sai về Hbase

- ☐ a.Hbase có hệ thuộc vào các dịch vụ cung cấp bởi HDFS
- ☐ b.Hbase có lệ thuộc vào các dịch vụ cung cấp bởi Zookeeper
- ☒ c.Hbase không hỗ trợ versioning
- ☒ d.Hbase hỗ trợ truy vấn dạng SQL

Question 30: Hệ thống nào cho phép đọc ghi dữ liệu tại vị trí ngẫu nhiên, thời gian thực tới hàng terabyte dữ liệu

- ☐ a.Flume
- ☐ b.Pig
- ☐ c.HDFS
- ☒ d.Hbase



Hadoop Ecosystem

- HDFS: Lưu trữ
- MapReduce: Framework xử lý dữ liệu (Coordination)
- YARN: Quản lý tài nguyên
- Spark: Xử lý dữ liệu trong RAM
- Pig, Hive: Xử lý dữ liệu query-based
- Mahout, MLlib: Machine learning
- Solar, lucence: Search và index
- Mesos, Zookeeper: Quản lý cụm (Maintain availability)
- Oozie: Orchestration (Giống airflow)
- Flume, Sqoop: Vận chuyển dữ liệu

Question 31: Đây là đặc điểm của RDD (Resilient distributed dataset) của Spark?

- ☒ a. ~~Người lập trình có thể quyết định số các phân mảnh của mỗi RDD~~
- ☐ b. Người sử dụng không thể quyết định số các phân mảnh của mỗi RDD
- ☒ c. ~~Được chia thành các phân mảnh (partition)~~
- ☒ d. Có khả năng chịu lỗi



Đặc điểm của RDD

- Bất biến (Chỉ đọc không sửa)
- Người dùng quyết định cách phân mảnh
- Mọi tính toán chia thành action hoặc transformation
- Chống lỗi
- Có khả năng lưu trữ trong cache
- Có tùy chọn lưu trong đĩa hoặc RAM

Question 32: Đây 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 (Debate)



Các transformation của RDD được tạo thành DAG

Question 33: Ưu điểm của kiến trúc SAN (Storage area network)?

- ☒ a. Hiệu năng, băng thông tốt hơn với NAS.
- ☒ b. Quản trị dễ dàng hơn so với NAS.
- ☐ c. Máy khách có thể kết nối tới SAN bằng đường truyền Ethernet thông thường (Chuẩn kết nối TCP/IP).

Question 34: Đây là kỹ thuật có thể được dùng để thích nghi các giải thuật học máy cho dữ liệu lớn?

- ☐ a. (2) Song song hoá trên Mapreduce hay Spark
- ☐ b. (3) Các kiến trúc mới xử lý luồng liên tục như mini-batch, complex event processing
- ☒ c. Tất cả các ý (1), (2), (3)

- ☐ d.(1) Sub-sampling, principal component analysis, feature extraction và feature selection
- ☐ e.Các ý (2) và (3)

Question 35: Phát biểu nào sau đây sai về Kafka?

- ☒ a.Kafka bảo đảm thứ tự của các message với mỗi topics.
- ☐ b.Message sau khi được tiêu thụ (consume) thì không bị xoá.
- ☐ c.Partition được nhân bản ra nhiều brokers.
- ☐ d.Các topic gồm nhiều partition

Question 36: Phát biểu nào sau đây sai về Kafka?

- ☐ a.Tất cả các thao tác ghi, đọc được xử lý bởi leader, follower làm theo leader.
- ☒ b.Mỗi partition có 1 leader và nhiều followers.
- ☐ c.Nếu leader bị lỗi, 1 follower sẽ thay thế trở thành leader mới

Question 37: Thế nào là UNIX semantic?

- ☐ a.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.
- ☒ 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.Tệp tin là chỉ đọc, không cho phép cập nhật và ghi đè. Mọi tiến trình đều có thể đọc tệp tin đồng thời.

Question 38: 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

Question 39: Đây là đặc điểm của Spark streaming?

- ☐ a. Không thể thực hiện các truy vấn SQL
- ☒ b. Có thể nhận đầu vào là các luồng dữ liệu từ Kafka
- ☒ c. Có thể nhận đầu vào là các tệp tin trên HDFS



Question 40: 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. Giao diện truy vấn đơn giản hơn so với CSQL quan hệ truyền thống
- ☒ c. Nhân bản (Replication)