CÂU HỎI VẤN ĐÁP

MÔN: XỬ LÝ DỮ LIỆU LỚN

Đề thi cho mỗi sinh viên là tổ hợp của một câu ở Phần I và một câu ở Phần II.

I. LÝ THUYẾT

1. Nêu các đặc trưng cơ bản của dữ liệu lớn. Với mỗi đặc trưng hãy nêu ví dụ thực tế để minh họa.
2. Nêu các lợi ích của xử lý dữ liệu lớn. Nêu ví dụ thực tế minh họa.
3. Nêu các thách thức khi xử lý dữ liệu lớn.
4. Trình bày các cách tiếp cận chính hiện nay để xử lý dữ liệu lớn.
5. Trình bày vắn tắt các đặc điểm chính của Hadoop. Nêu các ưu điểm, nhược điểm chính của Hadoop.
6. Trình bày vắn tắt các thành phần chính của Hadoop.
7. Mô tả mô hình lập trình MapReduce. Trình bày vắn tắt luồng xử lý dữ liệu trong MapReduce.
8. Trình bày các đặc điểm chính của nền tảng Apache Spark.
9. Trình bày các điểm khác biệt chính trong mô hình xử lý của Spark so với MapReduce.
10. Trình bày các thành phần chính của nền tảng Apache Spark.

II. THỰC HÀNH

1. Cho một tập dữ liệu văn bản. Hãy xây dựng thuật toán MapReduce thực hiện đếm số lần xuất hiện của các từ trong tập dữ liệu. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Java/Python).
2. Cho một tập dữ liệu văn bản. Hãy xây dựng thuật toán MapReduce thực hiện đếm tổng số từ trong tập dữ liệu. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Java/Python).
3. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tìm nhiệt độ cao nhất ghi nhận được theo từng năm. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Python/Java).
4. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tìm nhiệt độ thấp nhất ghi nhận được theo từng năm. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Python/Java).
5. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tính nhiệt độ trung bình theo từng năm. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Python/Java).
6. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tìm ngày, tháng có nhiệt độ cao nhất của mỗi năm. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Python/Java).
7. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tìm tọa độ địa lý (vĩ độ, kinh độ) của vị trí ghi nhận nhiệt độ cao nhất của mỗi năm. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Python/Java).
8. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tìm độ cao (tính bằng mét) của vị trí ghi nhận nhiệt độ cao nhất của mỗi năm. Cài đặt thuật toán này với Hadoop MapReduce bằng ngôn ngữ tùy chọn (Python/Java).

9. Cho tập dữ liệu thời tiết NCDC (đã tìm hiểu trong quá trình học). Hãy xây dựng thuật toán MapReduce tìm nhiệt độ cao nhất ghi nhận được mỗi năm. Cài đặt thuật toán này với PySpark.
10. Cho tập dữ liệu văn bản. Hãy xây dựng thuật toán MapReduce tìm n từ xuất hiện nhiều nhất. Cài đặt thuật toán này với PySpark.
