TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

HƯỚNG DẪN THỰC HIỆN
NHẬP MÔN DỮ LIỆU LỚN (BDES333877)

---

1. Hình thức thực hiện

- Nhóm: 03 – 05 sinh viên (số lẻ).
- Yêu cầu tổng hợp quá trình tìm hiểu và thực hiện thông qua trình bày quyển báo cáo đồ án (theo mẫu).
- Lưu trữ toàn bộ source-code, nguồn tài liệu tham khảo… lên github, gitlab,… hoặc các trang quản lý mã nguồn tương tự (tracking đóng góp thông qua submit code).
- Phân chia công việc rõ ràng (không tính thời gian làm báo cáo, slides, định dạng báo cáo – video,… hoặc các công việc không liên quan đến nội dung môn học)
- Khuyến khích quay video trình bày các bước thực hiện đồ án (theo hướng dẫn, mẫu).

2. Nội dung thực hiện

2.1. Cài đặt môi trường phát triển

- Hệ điều hành Ubuntu Server (**phiên bản mới nhất**).
- Apache Hadoop (phiên bản tương thích).
- Các công cụ trong Eco System Hadoop kèm phiên bản tương thích sử dụng.

Lưu ý: sinh viên cần cung cấp chi tiết thông tin cấu hình máy tính, hệ thống đang sử dụng: loại máy tính, dung lượng RAM, thông số CPU, thông số vi xử lý đồ họa (nếu có), phiên bản hệ điều hành, phiên bản ứng dụng. Trường hợp sử dụng máy ảo thì thêm thông số cấu hình máy ảo.

2.2. Yêu cầu thực hiện

- Thực hiện việc thu thập dữ liệu theo chủ đề (tự chọn) từ nhiều nguồn khác nhau (tối thiểu 2), đòi hỏi khối lượng dữ liệu thu thập phải đủ lớn theo phương pháp thủ công, tự động hoặc cả hai.
- Lưu trữ dữ liệu vào các DBMS (tự chọn), ví dụ: MySQL, MongoDB, PostgreSQL, Apache Cassandra, Google Cloud Datastore, Amazon Redshift, Azure Synapse Analytics, Amazon DynamoDB,...
- Cài đặt các ứng dụng Hadoop Eco System trên các môi trường khác nhau (Windows, MacOS, Docker).
- Tìm hiểu việc làm sạch dữ liệu (sửa chữa, loại bỏ dữ liệu không chính xác hoặc được định dạng không chính xác, trùng lặp dữ liệu, gán nhãn sai,…).
- Sử dụng các công cụ trong hệ sinh thái Hadoop vào việc truy vấn, lưu trữ, thao tác (CRUD), rút trích và tìm kiếm dữ liệu (tự đề xuất).
- Xây dựng giao diện tương tác cho hệ thống (optional).
- Trực quan hóa dữ liệu thu thập (optional).
- Xây dựng các chương trình MapReduce trong việc xử lý dữ liệu (tự đề xuất).
- Tất cả quá trình cài đặt, cấu hình cần trình bày chi tiết từng bước (step-by-step) trong báo cáo đồ án, **điểm cộng** khi nhóm thực hiện **thêm** công việc quay phim hướng dẫn thực hiện.

2.3. Báo cáo

- Nhóm sinh viên được yêu cầu trình bày thông qua quyển báo cáo đồ án môn học (theo mẫu).
- Soạn slides thuyết trình, trình bày những nội dung tìm hiểu và kết quả đạt được (10 – 15 slides).
- Có bảng phân công công việc thực hiện của các thành viên trong nhóm.

3. Đánh giá

- Báo cáo thuyết trình trước lớp (không bắt buộc, nhóm đăng ký trước ưu tiên → **điểm cộng**)
- Vấn đáp cá nhân từng thành viên trong nhóm dựa trên bảng phân công công việc, git (cuối kỳ) (bắt buộc).

4. Tiến độ thực hiện

- Báo cáo tuần 14 – 15.
- Nộp báo cáo đồ án môn học bản mềm (sẽ thông báo sau) tuần 16.

5. Hướng dẫn nộp đồ án

Nhóm tạo thư mục được đặt tên có cấu trúc sau:

<Mã lớp>_<STT nhóm>_<Tên đề tài>

Ví dụ: **06_01_NghienCuuXayDungHeThongABC**

Trong thư mục tổ chức như sau:

- Thư mục "**source-code**": chứa toàn bộ source code chương trình mà nhóm đã phát triển.
- Thư mục "**reports**" : báo cáo (*.docx và *.pdf), slides, bảng tự chấm, bảng phân công và video (nếu có).
- Thư mục "**dataset**" : chứa dữ liệu được sử dụng trong chương trình.
- Thư mục "**refs**" : chứa danh sách các tài liệu tham khảo (nếu có).
- Thư mục "**libs**" : danh sách các phần mềm, thư viện,… có liên quan (nếu có).
- Tập tin "**readme.txt**" : chứa thông tin có cấu trúc như sau:

```
----- Thông tin đề tài ----------------------
STT: ...
Tên đề tài: ...
Lớp học phần: <mã_học_phần>_xx
Năm học: HKx/20xx-20xx

----- Thông tin nhóm -----------------------
1. Họ tên sinh viên trưởng nhóm (mã số sinh viên trưởng nhóm) – SĐT – Email cá nhân
2. Họ tên sinh viên 2 (mã số sinh viên 2)
3. Họ tên sinh viên 3 (mã số sinh viên 3)
4. Họ tên sinh viên 4 (mã số sinh viên 4)
5. Họ tên sinh viên 5 (mã số sinh viên 5)
```

Sau khi thực hiện, nén thư mục với định dạng ***.zip** và nộp theo yêu cầu.

--- HẾT ---