IE405.E31
Công nghệ phân tích dữ liệu lớn

Huỳnh Công Bằng

GVHD: ThS. Nguyễn Hồ Duy Trí

Tp. Hồ Chí Minh, Ngày 22 tháng 7 năm 2025

Mục lục

LỜI MỞ ĐẦU

- Tuân thủ pháp luật trước khi học.
- Học cách tư duy và sơ đồ phát triển.
- Cần thực hành nhiều.
- Các công cụ nguyên thủy được viết trên Scala (ngôn ngữ hướng hàm).
- Mỗi lớp thực hành 2 buổi, sau đó báo cáo cuối kỳ (trong 4 buổi).

Chương 1

Giới thiệu

1.1 Mô tả môn học

- **Kiến thức cần thiết**: Toán rời rạc và toán thống kê.
- **Kiến thức được học**: Nền tảng dữ liệu lớn, mô hình và công cụ xử lý.
- **Kiến thức tự học:** Văn hóa doanh nghiệp.
- **Ngôn ngữ thực hành:** Python.

1.2 Mục tiêu

- **Hiểu và trình bày được các khái niệm cơ bản.**
- Tổ chức và lưu trữ được các nền tảng lưu trữ dữ liệu lớn.
- Áp dụng giải thuật Data mining, Machine learning, AI,... vào công việc.
- Tìm hiểu hệ thống và hệ sinh thái công ty lựa chọn.

1.3 Nội dung môn học

- Tổng quan: Lý do học dữ liệu lớn.
- Mô hình Mapreduce: Là mô hình ban đầu, tiền đề để phát triển.
- Framework Apache Spark: học kiến thức nền tảng về công cụ xử lý dữ liệu lớn.
- Hệ sinh thái dữ liệu lớn: vừa là đồ án cuối kỳ vừa là seminar.

1.4 Phân bố thời lượng

- Buổi 1: Giới thiệu
- Buổi 2: Giới thiệu (tiếp)
- Buổi 3: MapReduce
- Buổi 4: MapReduce (tiếp)
- Buổi 5: Apache Spark

- Buổi 6: Apache Spark (tiếp)
- Buổi 7: Thực hành 1 (lớp 1): Cài đặt Apache Spark và dùng thử
- Buổi 8: Thực hành 1 (lớp 2): Cài đặt Apache Spark và dùng thử
- Buổi 9: Thực hành 2 (lớp 1): Lập trình nâng cao với Apache Spark
- Buổi 10: Thực hành 2 (lớp 2): Lập trình nâng cao với Apache Spark
- Buổi 11: Thuyết trình: Hệ sinh thái dữ liệu lớn
- Buổi 12: Thuyết trình: Hệ sinh thái dữ liệu lớn (tiếp)
- Buổi 13: Thuyết trình: Hệ sinh thái dữ liệu lớn (tiếp)
- Buổi 14: Thuyết trình: (tiếp) + ôn tập

1.5 Đánh giá

Thực hành 20%, Semina + Bài tập 30%, Thi cuối ký 50%

1.6 Seminar

- Tìm hiểu một hệ quản trị dữ liệu lớn
- Tìm hiểu hệ thống xử lý bài toán dữ liệu lớn. (Real-time)

Chương 2

Tổng quan

2.1 Giới thiệu

2.1.1 Bối cảnh

Hãy mua trang thiết bị từ Samsung Smart Home để sống dễ hơn (tự nấu ăn tốn 3 tiếng với món khó thì nay chỉ tốn 1 tiếng, còn lại tự động hết), chưa kể đến rửa chén và dọn dẹp nhà cửa. Việc gì khó, có Samsung AI lo! Và dữ liệu lớn đã phát triển thành rất nhiều nhánh nhỏ như:

- Tính toán song song
- Lưu trữ khối lượng dữ liệu khổng lồ
- Phân tán dữ liệu
- Kết nối mạng tốc độ cao
- Máy tính hiệu suất cao
- Quản lý tác vụ và luồng xử lý
- Phân tích và khai thác dữ liệu
- Thu thập dữ liệu
- Học máy
- Trực quan hóa dữ liệu
- Đều là những công nghệ rất cũ

2.1.2 Quá trình phát triển

2.1.3 Dữ liệu hiện tại

2.2 Định nghĩa

2.3 Đặc điểm

2.4 Phân loại

2.5 Quy trình xử lý

2.6 Ứng dụng và xu hướng

2.7 Thách thức và cơ hội

2.8 Q & A