ĐỀ XUẤT ĐỀ TÀI ĐỒ ÁN MÔN HỌC

LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN

Xây dựng data pipeline để phân tích dữ liệu giải bóng bầu dục mỹ NFL

Giảng viên: TS. Nguyễn Bình Minh

Sinh Viên: Nhóm 5

Môn học: IT4931 - Lưu trữ và xử lý dữ liệu lớn

Lớp học: 136051

I. Danh sách thành viên

- Hữu Tường Tú - 20194395 - CTTN KHMT K64

- Nguyễn Phúc Tân - 20194163 - CTTN KHMT K64

- Nguyễn Thành Phong - 20192016 - CTTN KHMT K64

- Trần Tiến Bằng - 20193988 - CTTN KHMT K64

II. Giới thiệu đề tài

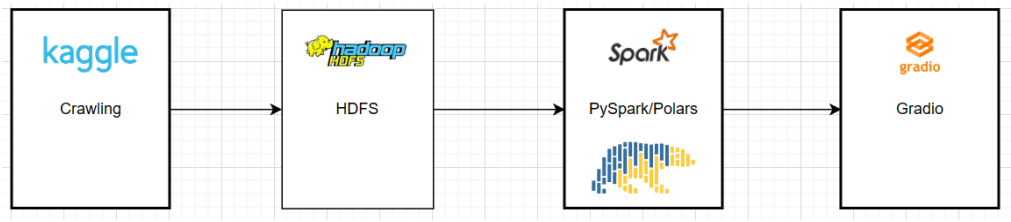
1. Mô tả bài toán

Bóng bầu dục Mỹ hay còn gọi là bóng đá kiểu Mỹ *(American football* hoặc *Gridiron football)* là một môn thể thao phổ biến tại nhiều nơi trên thế giới. Mỗi trận đấu bóng bầu dục sẽ có các thống kê của từng đội hay từng cầu thủ. Qua các mùa giải một số lượng lớn trận đấu sẽ được diễn ra và việc xử lý thông tin của các trận đấu có thể mang lại những thông tin về các chiến thuật của các đội hoặc các chỉ số thông tin của các vận động viên. Chính vì vậy, chúng em sẽ thu thập, lưu trữ, xử lý và phân tích dữ liệu lớn cần thiết để có thể đưa ra những phân tích cho đề tài.

2. Mục tiêu và phạm vi của đề tài

Thông qua khảo sát và phân tích, đề tài của chúng em sẽ lấy dữ liệu trận đấu bóng bầu dục NFL năm 2017 đến năm 2019. Sau đó chúng em sẽ xử lý và thống kê những chỉ số như độ tuổi, chiều cao, cân nặng các cầu thủ, các nơi xuất ra được nhiều cầu thủ nhất, các chiến thuật hay sử dụng, các vị trí cầu thủ hay xuất phát khi ở trên sân... từ đó để đánh giá những phương án và xu hướng, chiến thuật các đội.

3. Project Pipeline



Bước 1: Thu thập dữ liệu.

- Dữ liệu sẽ được thu thập từ trên kaggle [NFL Big Data Bowl | Kaggle]

Bước 2: Lưu trữ dữ liệu.

- Lưu trữ dữ liệu đã trên HDFS

Bước 3: Truy vấn dữ liệu:

- Sử dụng pyspark/polars để truy vấn dữ liệu và so sánh hiệu năng.

Bước 4: Mô phỏng dữ liệu:

- Kết hợp matplotlib và gradio để mô phỏng dữ liệu
- Phân tích dữ liệu từ trên những đồ thị đã mô phỏng