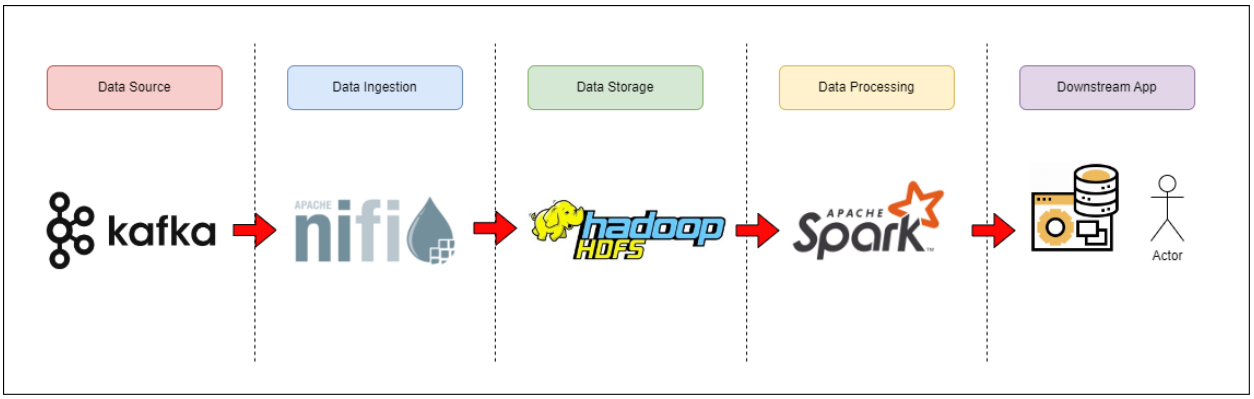
Assignment – DE VDT2024

1. Bài toán

Xây dựng 1 nền tảng dữ liệu đơn giản bằng 1 số công nghệ được đào tạo trong chương trình.



2. Yêu cầu thực hiện

a. Viết chương trình đẩy dữ liệu lên Kafka Topic.
- Dữ liệu đọc từng dòng từ file "log_action.csv.
- Đẩy dữ liệu lên Kafka Topic "vdt2024" mỗi giây, bao gồm các trường:

Trường	Kiểu dữ liệu	Mô tả
student_code	Number	Là số thự tự của sinh viên trong lớp DE
activity	String	Hoạt động, lấy ngẫu nhiên trong danh sách: ["read", "write", "execute"]
numberOfFile	Number	Số file xử lý
timestamp	String	Thời điểm bản ghi được tạo

b. Triển khai Nifi, kéo dữ liệu từ Kafka Topic "vdt2024", xử lý và lưu dữ liệu xuống HDFS với đường dẫn: "/raw_zone/fact/activity". Lưu dữ liệu dưới dạng parquet.

c. Lưu trữ file "danh_sach_sv_de.csv" xuống HDFS

d. Viết chương trình xử lý dữ liệu lưu trữ dưới HDFS, sử dụng Apache Spark.

Coding xử lý dữ liệu ở trên, yêu cầu: Đưa ra tổng số file được tương tác hàng ngày theo mỗi loại activity mà sinh viên đó thực hiện. Lưu ra 1 file ouput.

File Output:

- Tên file: Tên_sinh_viên.csv
- Định dạng: CSV
- Schema: date, student_code, student_name, activity, totalFile

Ví dụ output:

20240519, 1, Mai Đức An, read, 5

20240519, 1, Mai Đức An, execute, 2

20240520, 1, Mai Đức An, read, 3

3. Yêu cầu đầu ra

1 bản báo cáo 8-10 trang (kèm link source code trên github), mô tả:

- Cách triển khai các thành phần trong nền tảng: Mô tả cách thức triển khai, thực hiện các bước.
- Upload kết quả file Output sau khi chạy xử lý ở bước 3.