KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện Thoại: (08) 38.354.266 - Fax:(08) 38.350.096

CSC17106 – XỬ LÝ PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN

HƯỚNG DẪN THỰC HÀNH

ĐỒ ÁN MÔN HỌC

I. Thông tin chung

Mã số:	HD10
Thời lượng dự kiến:	90 tiếng
Deadline nộp bài:	-
Hình thức:	Vấn đáp
Hình thức nộp bài:	Moodle
GV phụ trách:	Phạm Minh Tú
Thông tin liên lạc với GV:	pmtu@fit.hcmus.edu.vn

II. Mô tả bài toán

Một công ty tài chính muốn xây dựng hệ thống quản lý dữ liệu giao dịch thông qua thẻ tín dụng (credit card), dữ liệu được phát sinh từ các máy POS đặt tại các cửa hàng mua sắm, nhà hàng, bất cứ nơi nào thanh toán không dùng tiền mặt. Công ty muốn xây dựng hệ thống xử lý dữ liệu theo thời gian thực, khi một giao dịch được phát sinh, dữ liệu được gửi đến hệ thống, tiến hành kiểm tra dữ liệu có lỗi hay không? Is Fraud = Yes xác định lỗi và giao dịch này xem như không thành công, không cần xử lý tiếp. Khi một giao dịch thành công thì tiến hành **lưu trữ** các thông tin Credit Card, ngày giao dịch theo định dạng: dd/mm/yyyy, thời gian theo định dạng: hh:mm:ss, Merchant name (nơi xảy ra giao dịch), Merchant City (thành phố nơi giao dịch), Số tiền chuyển sang VNĐ, theo tỉ giá được cập nhật mỗi ngày. Cuối ngày, tất cả giao dịch được được thống kê như sau: Cho biết tổng số lượng giá trị từng merchant name, đếm số lượng giao dịch mỗi merchant name, tất cả thống kê theo ngày, tháng và năm. Tất cả thông tin thống kê này được trực quan hóa qua công cụ hoặc hệ thống chuyên biệt.

III. Các tình huống giả định

Hãy giả định rằng giao dịch được phát sinh mỗi khi người dùng dùng credit card quẹt trên các máy POS tại các cửa hàng mua sắm, nhà hàng,…. Mỗi giao dịch này được gửi qua hệ thống kafka theo thời gian thực. Sinh viên dùng Kafka để mô phỏng từng giao dịch được phát sinh với các thông tin được cho trước dạng csv. (tập tin sẽ được đính kèm)

Cấu trúc thông tin như sau:

User,Card,Year,Month,Day,Time,Amount,Use Chip,Merchant Name,Merchant City,Merchant State,Zip,MCC,Errors?,Is Fraud?

Kafka sẽ đọc từng dòng csv và gửi qua topic được định nghĩa trước để giả lập một giao dịch được phát sinh từ máy POS.

```
User,Card,Year,Month,Day,Time,Amount,Use Chip,Merchant Name,Merchant City,Merchant State,Zip,MCC,Errors?,Is Fraud?
0,0,2002,9,1,06:21,$134.09,Swipe Transaction,3527213246127876953,La Verne,CA,91750.0,5300,,No
0,0,2002,9,1,06:42,$38.48,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,2,06:22,$120.34,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,2,17:45,$128.95,Swipe Transaction,3414527459579106770,Monterey Park,CA,91754.0,5651,,No
0,0,2002,9,3,06:23,$104.71,Swipe Transaction,5817218446178736267,La Verne,CA,91750.0,5912,,No
0,0,2002,9,3,13:53,$86.19,Swipe Transaction,-7146670748125200898,Monterey Park,CA,91755.0,5970,,No
0,0,2002,9,4,05:51,$93.84,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,4,06:09,$123.50,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,5,06:14,$61.72,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,5,09:35,$57.10,Swipe Transaction,4055257078481058705,La Verne,CA,91750.0,7538,,No
0,0,2002,9,5,20:18,$76.07,Swipe Transaction,-4500542936415012428,La Verne,CA,91750.0,5814,,No
0,0,2002,9,5,20:41,$53.91,Online Transaction,-9092677072201095172,ONLINE,,,4900,,No
0,0,2002,9,6,06:16,$110.37,Swipe Transaction,2027553650310142703,Mira Loma,CA,91752.0,5541,,No
0,0,2002,9,7,06:16,$117.05,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,7,06:34,$45.30,Swipe Transaction,-5475680618560174533,Monterey Park,CA,91755.0,5942,,No
0,0,2002,9,7,09:39,$29.34,Swipe Transaction,4055257078481058705,La Verne,CA,91750.0,7538,,No
0,0,2002,9,8,06:10,$147.45,Swipe Transaction,-34551508091458520,La Verne,CA,91750.0,5912,,No
0,0,2002,9,8,06:38,$27.75,Swipe Transaction,4060646732831064559,La Verne,CA,91750.0,5411,,No
0,0,2002,9,8,13:48,$76.57,Swipe Transaction,-727612092139916043,Monterey Park,CA,91754.0,5411,,No
0,0,2002,9,8,22:01,$22.56,Swipe Transaction,-6733168469687845480,Mira Loma,CA,91752.0,7832,,No
0,0,2002,9,9,06:54,$37.50,Swipe Transaction,4060646732831064559,La Verne,CA,91750.0,5411,,No
0,0,2002,9,9,09:40,$65.50,Swipe Transaction,-3345936507911876459,La Verne,CA,91750.0,7538,Technical Glitch,No
0,0,2002,9,9,13:19,$56.42,Swipe Transaction,3189517333335617109,La Verne,CA,91750.0,5311,,No
0,0,2002,9,9,13:31,$2.71,Swipe Transaction,4060646732831064559,La Verne,CA,91750.0,5411,,No
```

IV. Yêu cầu tối thiểu đồ án

Công nghệ:

1. Sử dụng kafka để đọc dữ liệu csv từng dòng và gửi thông tin này đến topic định nghĩa trước theo chu kì thời gian ngẫu nhiên trong phạm vi từ 1s đến 5s.

2. Sử dụng spark streaming để đọc dữ liệu từ kafka theo thời gian thực, nghĩa là bất cứ thông tin nào từ kafka được xử lý tức thì, các xử lý bao gồm lọc dữ liệu, biến đổi thông tin, tính toán dữ liệu bao gồm lấy tỉ giá mới nhất từ các web site chính thống như vietcombank, dùng Web Scraping và API (trường hợp API lỗi, không ổn định thì có phương án dự phòng là Web Scraping).

3. Sử dụng Hadoop để lưu trữ các thông tin được xử lý từ Spark và là nơi lưu trữ thông tin được xử lý để có thể trực quan hóa dữ liệu và thống kê ở giai đoạn sau.

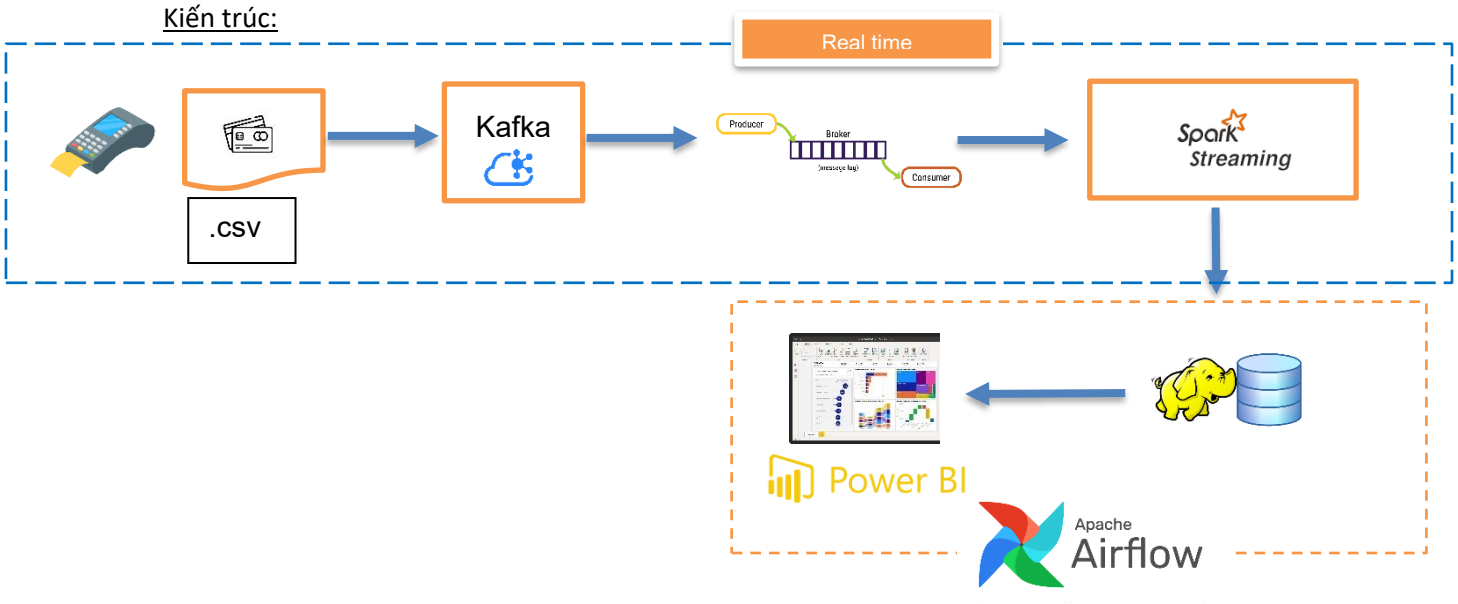
4. Sử dụng Power BI để đọc dữ liệu từ Hadoop (dạng csv), thống kê dữ liệu theo mô tả bài toán và hiển thị dữ liệu một cách trực quan.

5. Sử dụng Air Flow để lên lịch quá trình đọc và hiển thị dữ liệu từ Power PI sao cho dữ liệu luôn được update mỗi ngày.

Phân tích thời gian thực (*):

1. Thời điểm nào trong ngày có nhiều giao dịch nhất? Có khung giờ nào giao dịch bất thường không?
2. Thành phố nào có tổng giá trị giao dịch cao nhất? Có liên hệ với dân số hoặc vị trí không?
3. Merchant nào có số lượng hoặc giá trị giao dịch cao nhất?
4. Thành phố hoặc merchant nào có tỷ lệ fraud cao bất thường?
5. Người dùng nào có nhiều giao dịch liên tiếp trong thời gian ngắn?
6. Giao dịch có giá trị lớn thường xảy ra vào thời điểm nào? Ở đâu?
7. Có xu hướng nào trong các giao dịch bị fraud không? (giờ, merchant, city,...)
8. Có sự khác biệt nào giữa giao dịch ngày thường và cuối tuần?
9. Có người dùng nào bị nhiều lỗi hoặc bị gắn cờ fraud nhiều hơn mức trung bình?
10. Từ các phân tích trên, hãy đề xuất cải tiến cho hệ thống để giảm gian lận hoặc tối ưu vận hành.

Kiến trúc:


Real time
Kafka
.csv
Producer
Broker
Consumer
Spark
Streaming
Power BI
Apache
Airflow