Real-Time Iowa Liquor Sales Analysis

Nguyễn Minh Quân[19522082], Hồ Nhất Thống[19522300], Phạm Hoàng Thư[19522310], and Trần Thiên Thạch[19522182]

University of Information Technology – VNUHCM, Vietnam

Tóm tắt nội dung Ngữ cảnh của ứng dụng dựa trên hoạt động kho rượu lớn của bang Iowa. Khi khách hàng thực hiện thanh toán dữ liệu hóa đơn sẽ được lưu trữ tại flat file cụ thể là CSV. Hệ thống sẽ thu thập và xử lý dữ liệu các hóa đơn một cách realtime. Dữ liệu phân tích sẽ được sử dụng với 2 mục đích: một là lưu vào database (Postgres) cung cấp các dữ liệu đã được xử lý phù hợp cho các đội kinh doanh hoặc Business Inteligent xem báo cáo cũng như làm các phân tích đơn giản, hai là sử dụng để huấn luyện các mô hình trên bộ dữ liệu này theo hướng tiếp cận là sử dụng các thư viện Spark MLlib (Machine Learning) để dự đoán số lượng chai được phân phối ra ngoài thị trường.

Keywords: Real-Time · BigData · Streaming · Analysis

1 Giới thiệu

Rượu là một mặt hàng kinh doanh quan trọng và có giá trị kinh tế cao trên toàn cầu. Ngành sản xuất và tiêu thụ rượu đóng góp một phần không nhỏ vào nền kinh tế của nhiều quốc gia. Về mặt sản xuất, ngành công nghiệp rượu được coi là một ngành kinh doanh lớn, tạo ra nhiều việc làm cho các nông dân, nhà sản xuất nho, nhà sản xuất bia, nhà phân phối và các công ty kinh doanh rượu. Ngoài ra, việc sản xuất rượu cũng đóng góp vào nền kinh tế địa phương bằng cách tăng cường nguồn thu nhập cho những khu vực nông thôn và tăng cường khả năng xuất khẩu cho các quốc gia sản xuất rượu. Về mặt tiêu thụ, rượu là một mặt hàng xa xỉ và đắt đỏ, đóng góp một phần không nhỏ vào ngành du lịch và giải trí. Việc tiêu thụ rượu cũng đóng góp vào nền kinh tế của một số quốc gia nhờ thu thuế và tăng cường doanh thu từ việc bán rượu. Tuy nhiên, rượu cũng gây ra một số vấn đề kinh tế khác, chẳng hạn như tác động tiêu cực đến sức khỏe và xã hội. Việc tiêu thụ quá nhiều rượu có thể gây ra các vấn đề sức khỏe như ung thư, bệnh gan, bệnh tim mạch và bệnh thần kinh. Ngoài ra, rượu cũng có thể gây ra tai nạn giao thông và các vấn đề an ninh, gây thiệt hại đến kinh tế và xã hội.Vì vậy, việc kiểm soát và quản lý tiêu thụ rượu là rất quan trọng để giảm thiểu những tác động tiêu cực đến sức khỏe và xã hội, đồng thời tăng cường tác động tích cực của ngành rượu đối với nền kinh tế và xã hội. Bằng cách phân tích thời gian thực có thể giúp cơ quan chức năng kiểm soát và quản lý việc tiêu thụ rượu trên thị trường một cách trực quan và nhanh chóng.

Thay vì cố gắng sửa đổi các mô hình học máy để cải thiện hiệu suất dự báo tiêu thụ rượu trên thì trường, thì báo cáo này tập trung vào trình bày việc triển

khai ứng dụng các công cụ Big Data như Apache Kafka, Apache Spark vào việc xử lý dữ liệu lớn trong thời gian thực. Cụ thể hơn, dữ liệu về việc phân phối rượu ra thị trường sẽ được thu thập và phát trực tuyến qua Kafka. Dữ liệu phát trực tuyến sau đó được đưa vào mô hình máy học dự đoán đã được đào tạo được tích hợp vào Spark Structerd Streaming. Cuối cùng dữ liệu dự đoán sẽ được phát theo thời gian thực, cùng với các phân tích khác trên Power BI.

2 Bộ dữ liệu

2.1 Giới thiệu bộ dữ liệu

Iowa Liquor Sales là bộ dữ liệu chứa thông tin mua rượu mạnh của những người được cấp phép rượu cấp E của Iowa theo sản phẩm và ngày mua từ ngày 1 tháng 1 năm 2021 đến năm 2022. Bộ dữ liệu có thể được sử dụng để phân tích tổng doanh số bán rượu mạnh ở Iowa của từng sản phẩm ở cấp độ cửa hàng. Giấy phép rượu loại E, dành cho cửa hàng tạp hóa, cửa hàng rượu, cửa hàng tiện lợi, v.v., cho phép các cơ sở thương mại bán rượu để tiêu dùng ngoài cơ sở trong các thùng chứa ban đầu chưa mở. Bộ dữ liệu này chứa thông tin về tên, loại, giá cả, số lượng và địa điểm bán hàng của các thùng chứa riêng lẻ hoặc các gói thùng chứa đồ uống có cồn. Dữ liệu này có thể là một mẫu đại diện cho hoạt động bán rượu ở bang Iowa Hoa Kỳ và có thể được sử dụng để trả lời nhiều câu hỏi liên quan, chẳng hạn như: bao nhiêu rượu được bán và tiêu thụ ở Hoa Kỳ? Loai nào? Các nhãn hiệu và nhãn phổ biến nhất là gì ?

2.2 Xử lý dữ liệu

Bộ dữ liệu chứa gần 3 triệu dòng và 24 cột thuộc tính. Trong bài báo cáo này chúng tôi phân tích với quy mô là quận và thành phố, đồng thời các chỉ số tiêu thụ có đơn vị đo là chai và mililit(ml) nên sẽ loại bỏ các cột thuộc tính không cần thiết. Nhìn chung dữ liệu đã được làm sạch, các thuộc tính sẽ được định dạng lại kiểu dữ liệu và các giá trị null sẽ được thay thế.

3 Mô hình hệ thống

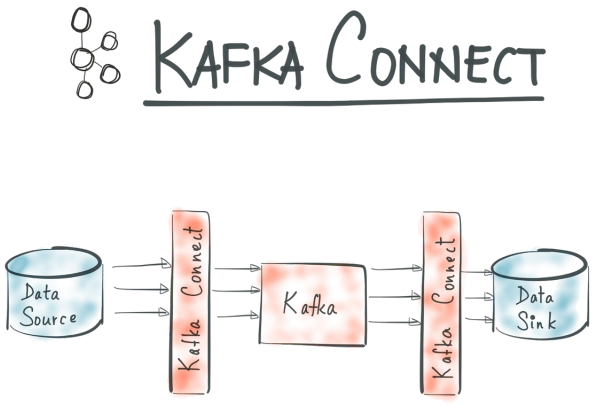
Kiến trúc của mô hình Real-time được minh họa trong Hình 1. Như có thể thấy trong hình, dữ liệu ban đầu được lấy từ file CSV do kho rượu Iowa cung cấp. Dữ liệu này được truyền phát bằng Kafka và được đưa vào các thành phần được tích hợp bên trong Spark Strucrted Streaming. Sau khi dữ liệu được xử lý và chọn dữ liệu bằng SQL sẽ được lưu vào database. Từ đó data team và business team sẽ sử dụng database này vào việc nghiên cứu và phân tích dữ liệu.

3.1 Frameworks dữ liệu lớn và docker

1.Kafka

là một công nghệ truyền dữ liệu phân tán (distributed messaging system) theo mô hình truyền thông public-subscribe, bên truyền dữ liệu được gọi là producer bên subscribe nhận dữ liệu theo các topic được gọi là consumer. Kafka có khả năng truyền một lượng lớn dữ liệu tuy nhiên trong trường hợp khi consumer chưa nhận, dữ liệu vẫn được lưu trữ sao lưu trên queue và cả trên ổ đĩa bảo đảm an toàn.

Kafka Connect một thành phần của Kafka, dùng để kết nối Kafka với các hệ thống khác như các database, file system, key-value store... Kafka Connect Cluster sẽ tách biệt với Kafka cluster với mục đích để có thể scale các connector bên trong nó.



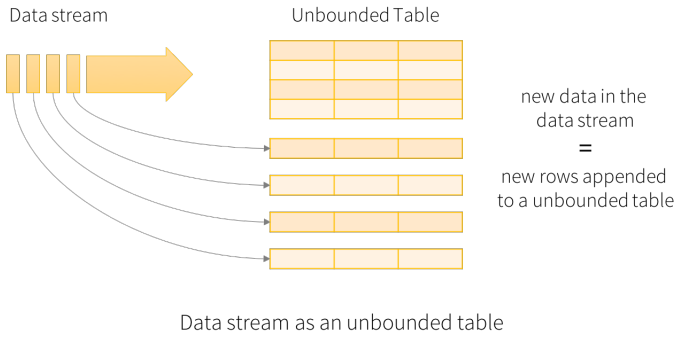
Hình 1. Kafka Connect

Kafka Connector được thiết kế để chạy trong Kafka Connect Cluster, thành phần này sẽ được sử dụng để đọc dữ liệu từ các nguồn khác vào kafka topic hoặc đọc dữ liệu từ kafka topic gửi đến các nguồn khác.

2. Apache Spark

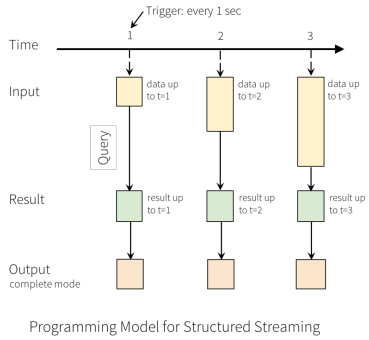
là một framework mã nguồn mở tính toán cụm. Tốc độ xử lý của Spark có được do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau. Đồng thời việc tính toán được thực hiện ở bộ nhớ trong (in-memories) hay thực hiện hoàn toàn trên RAM.Về bản chất Spark sẽ không xử lý dữ liệu streaming như hình thức của Apache Flink, mà spark sẽ xử lý dữ liệu theo từng micro batch và ta có thể config interval của từng batch sao cho phù hợp. Với việc mỗi micro-batch có thời gian rất nhỏ nên việc spark xử lý dữ dữ liệu gần như streaming.Truyền dữ liệu, là loại dữ liệu được tạo liên tục như nhận xét từ mạng xã hội, được chuyển đến các mô hình được tích hợp bên trong spark structured

streaming và xử lý ngay sau khi thời điểm nó đến để tạo ra kết quả trong thời gian thực.



Hình 2. Truyền dữ liệu vào bảng

Như hình trên ta có thể thấy dữ liệu streaming sẽ thêm vào một bảng không giới hạn và thời gian của mỗi micro-batch ta có thể tùy chỉnh được. Lấy ví dụ khi thời gian của mỗi micro-batch là 1s ta có thể hiểu spark streaming vận hành theo cách sau



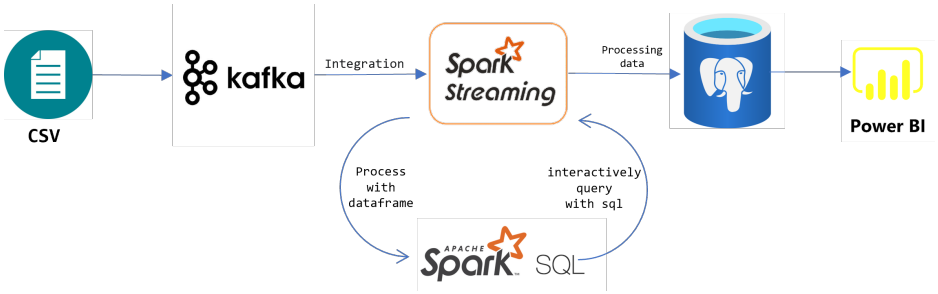
Hình 3. cách spark streaming vận hành

Sau khi đã có được kết quả của query thì Spark sẽ cần lưu trữ kết quả này vào một nơi lưu trữ nào đó theo 1 trong 3 chế độ sau: Complete: Spark sẽ lưu

lại toàn bộ kết quả xử lý được tính tới thời điểm gần nhất Update: Spark sẽ chỉ lưu lại các dữ liệu mới tính tại thời điểm gần nhất. Trong trường hợp không thể thay đổi được dữ liệu ở nơi lưu trữ thì các dữ liệu này sẽ được thêm vào như là một dữ liệu mới Append: Spark sẽ chỉ lưu lại các dữ liệu mới vào nơi lưu trữ, tính tại thời điểm gần nhất

3.Docker

Là một nền tảng cho developers và sysadmin để develop, deploy và run application với container. Nó cho phép tạo các môi trường độc lập và tách biệt để khởi chạy và phát triển ứng dụng và môi trường này được gọi là container. Trong bài báo cáo này docker được sử dụng để deploy zookeper, kafka, postgresql nhằm việc tạo môi trường mộ cách nhanh chóng và dễ dàng.



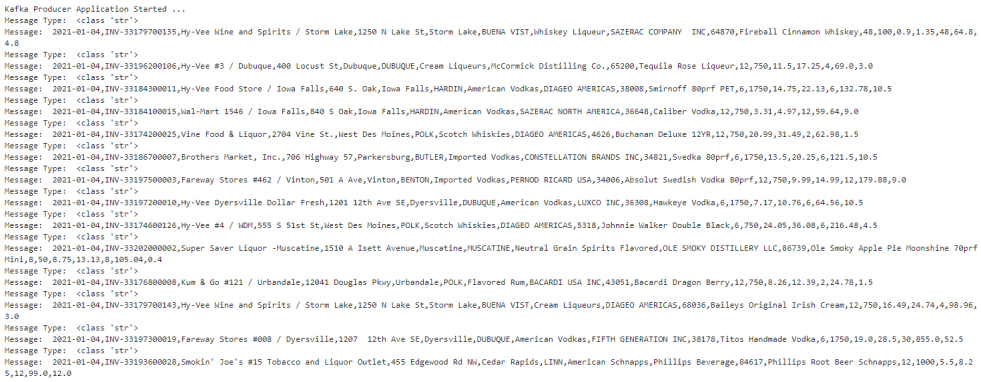
Hình 4. Tổng quan kiến trúc

3.2 Chi tiết về hệ thống

Sử dụng cổng 'bootstrap-server:29092' gửi dữ liệu đến máy chủ của Kafka sau khi thu thập chúng. Sau khi Dữ liệu đầu tiên được lưu trữ trong Kafka, nó được tích hợp với Spark và được phát trực tuyến tới Spark. sau đó dữ liệu sẽ được truy vấn bởi SparkSQL. Spark sẽ lưu dữ liệu vào PostgreSQL bằng method DF.WriteStream vì thế dữ liệu sẽ liên tục được ghi vào database. PowerBI sẽ lấy dữ liệu từ Postgres thông qua cổng host.docker.internal:5432 và thực hiện phân tích.

4 Kết quả

Khi kafka thu thập và lấy dữ liệu từ csv, sẽ nối 17 cột thuộc tính thành một string và gửi thông qua cổng bootstrap-server:29092 với topic đã được cài đặt sẵn.



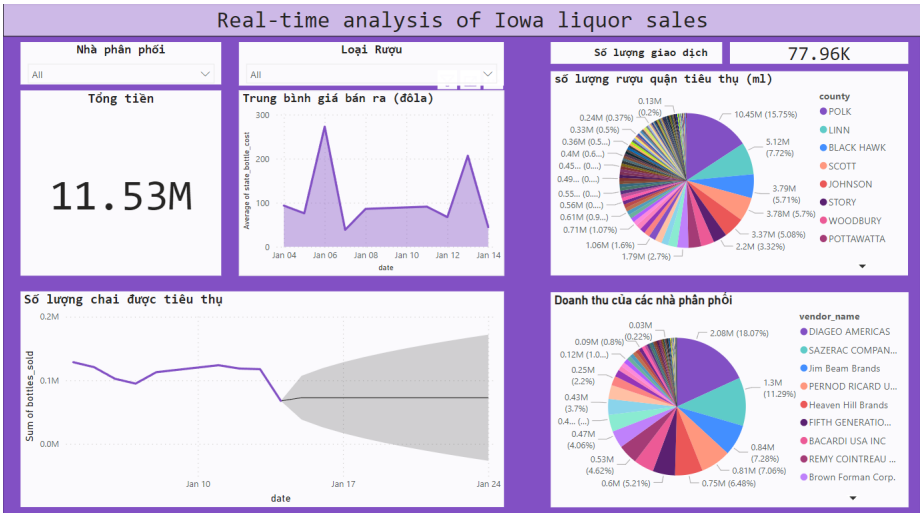
Hình 5. kafka gửi dữ lệu đến cổng

Spark đọc được dữ liệu từ cổng chuyển dữ liệu sang dataframe và định dạng lại kiểu dữ liệu, sử dụng sparkSQL để query dữ liệu khi triển khai WriteStream sau đí lưu dữ liệu vào Postgres.

	date timestamp without time zone	invoice_and_item_number text	store_name text	address text	city text	county text	category_name text	vendor_name text
1	2021-01-04 00:00:00	INV-33179700135	Hy-Vee Wine and Spirits / Storm Lake	1250 N Lake St	Storm Lake	BUENA VIST	Whiskey Liqueur	SAZERAC COMPANY INC
2	2021-01-04 00:00:00	INV-33196200106	Hy-Vee #3 / Dubuque	400 Locust St	Dubuque	DUBUQUE	Cream Liqueurs	McCormick Distilling Co.
3	2021-01-04 00:00:00	INV-33184300011	Hy-Vee Food Store / Iowa Falls	640 S. Oak	Iowa Falls	HARDIN	American Vodkas	DIAGEO AMERICAS
4	2021-01-04 00:00:00	INV-33184100015	Wal-Mart 1546 / Iowa Falls	840 S Oak	Iowa Falls	HARDIN	American Vodkas	SAZERAC NORTH AMERICA
5	2021-01-04 00:00:00	INV-33174200025	Vine Food & Liquor	2704 Vine St.	West Des Moines	POLK	Scotch Whiskies	DIAGEO AMERICAS
6	2021-01-04 00:00:00	INV-33186700007	Brothers Market	Inc.	706 Highway 57	Parkersburg	BUTLER	Imported Vodkas
7	2021-01-04 00:00:00	INV-33197500003	Fareway Stores #462 / Vinton	501 A Ave	Vinton	BENTON	Imported Vodkas	PERNOD RICARD USA
8	2021-01-04 00:00:00	INV-33197200010	Hy-Vee Dyersville Dollar Fresh	1201 12th Ave SE	Dyersville	DUBUQUE	American Vodkas	LUXCO INC
9	2021-01-04 00:00:00	INV-33174600126	Hy-Vee #4 / WDM	555 S 51st St	West Des Moines	POLK	Scotch Whiskies	DIAGEO AMERICAS
10	2021-01-04 00:00:00	INV-33202000002	Super Saver Liquor -Muscatine	1510 A Isett Avenue	Muscatine	MUSCATINE	Neutral Grain Spirits Flavored	OLE SMOKY DISTILLERY LLC
11	2021-01-04 00:00:00	INV-33176800008	Kum & Go #121 / Urbandale	12041 Douglas Pkwy	Urbandale	POLK	Flavored Rum	BACARDI USA INC

Hình 6. dữ liệu được ghi vào postgres

PowerBI kết nối với Postgres qua cổng host.docker.internal:5432, sau đó thực hiện phân tích cơ bản như: Tổng số tiền rượu, Số lượng giao dịch tại thời điểm đó, trung bình giá bán ra của mỗi sản phẩm, Số lượng rượu mà các quận tiêu thụ, doanh thu của các nhà phân phối, Số lượng chai được tiêu thụ và dự đoán khả năng tiêu thụ trong tương lai.



Hình 7. Phân tích thời gian thực

Tại thời điểm phân tích từ ngày 4/1/2021 đến ngày 14/1/2022, có thể thấy tổng doanh thu là 11,53 triệu đô do rượu mang lại. Tại quận Polk bang Iowa có tỉ lệ tiêu thụ lượng rượu tính trên đơn vị ml là 15,75% tỉ lệ cao nhất bang với hơn 10 triệu ml rượu được tiêu thụ Về phần dự đoán khả năng tiêu thụ, hiện tại thì sai số còn khá cao vì lượng dữ liệu đầu vào chưa nhiều để có thể dự đoán có phần chính xác cao hơn.

5 Tổng kết

Trong báo cáo này, chúng tôi đã tập trung vào việc xây dựng hệ thống phân tích thời gian thực với bộ dữ liệu iowa-liquor-sales dự trên Spark Structured Streaming và các công cụ khác. Kết quả đạt được là hệ thống phân tích thời gian thực có khả năng truyền, xử lý, trực quan hóa dữ liệu với thời gian thực. Hệ thống vẫn còn điểm yếu trong việc dự đoán do phụ thuộc vào bộ dữ liệu, đồng thời phần cứng không đủ đáp ứng nhu cầu. Trong tương lai tiếp tục giải quyết các bài toán còn sót lại n để hoàn thiện đề tài, cải thiện thêm bộ nhớ, các thiết bị, thiết lập tốt hơn giúp xử lí những bộ dữ liệu lớn hơn. Phát triển thêm các

chức năng mới như: phân cấp người quản lí, dự đoán xu hướng dựa trên máy học với sparkML.

Tài liệu

1. GABRIEL RAMOS, Updated Iowa Liquor Sales, Kaggle (2023).
2. Spark.apache.org,Structured Streaming + Kafka Integration Guide,document (2023).
3. Kafka.apache.org,Producer API,document(2023).
4. Nguyen Truong,Xây dựng mô hình Real-time Analytic,linkedin, (2023)