GỢI Ý DỰA TRÊN KỸ THUẬT PHÂN TÍCH CỤM SỬ DỤNG THUẬT TOÁN KMEANS

Nguyễn Tiến An - Hồ Tú Minh

MỤC LỤC

1.1 ĐỊNH NGHĨA

một lượng dữ liệu khổng lồ và phức tạp tới mức các công cụ quản lý dữ liệu truyền thống không có khả năng thu thập, quản lý và xử lý dữ liệu

1.2 CÁC LOẠI DỮ LIỆU

- Dữ liệu có cấu trúc
- Dữ liệu bán cấu trúc
- Dữ liệu phi cấu trúc

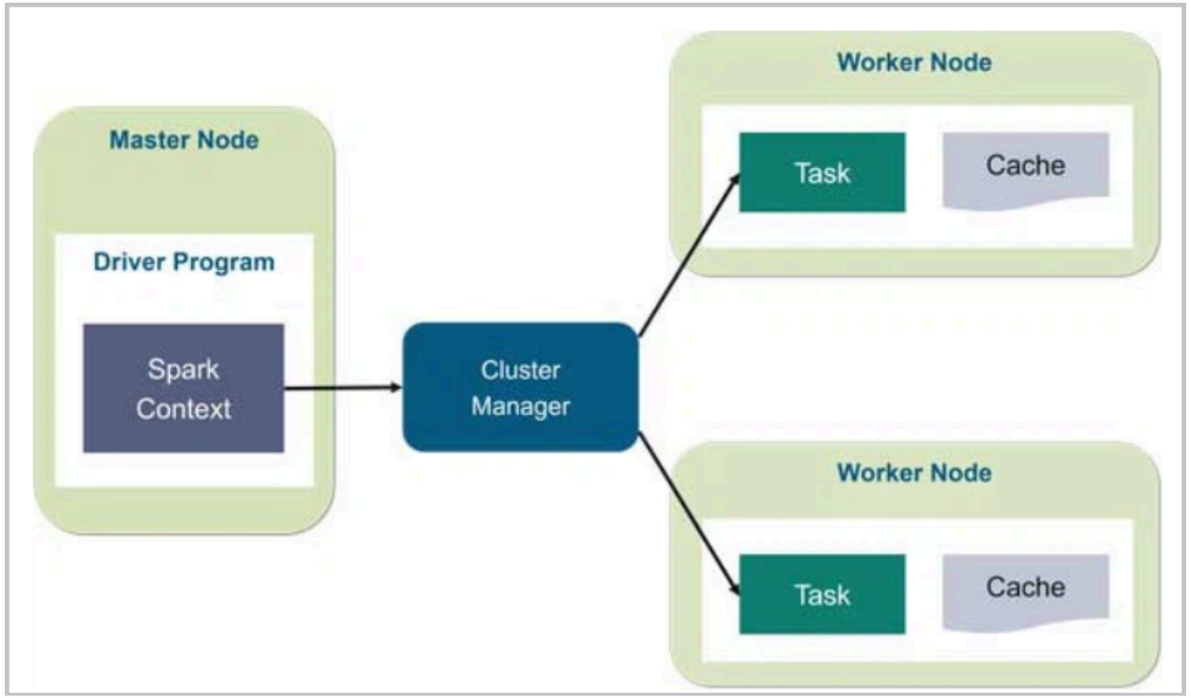
1.3 CÁC ĐẶC TRƯNG

- Khối lượng (Volume)
- Tốc độ (Velocity)
- Đa dạng (Variety)
- Độ chính xác (Veracity)
- Tính biến đổi (Variability)
- Giá trị (Value)

1.4 TỔNG QUAN VỀ APACHE SPARK

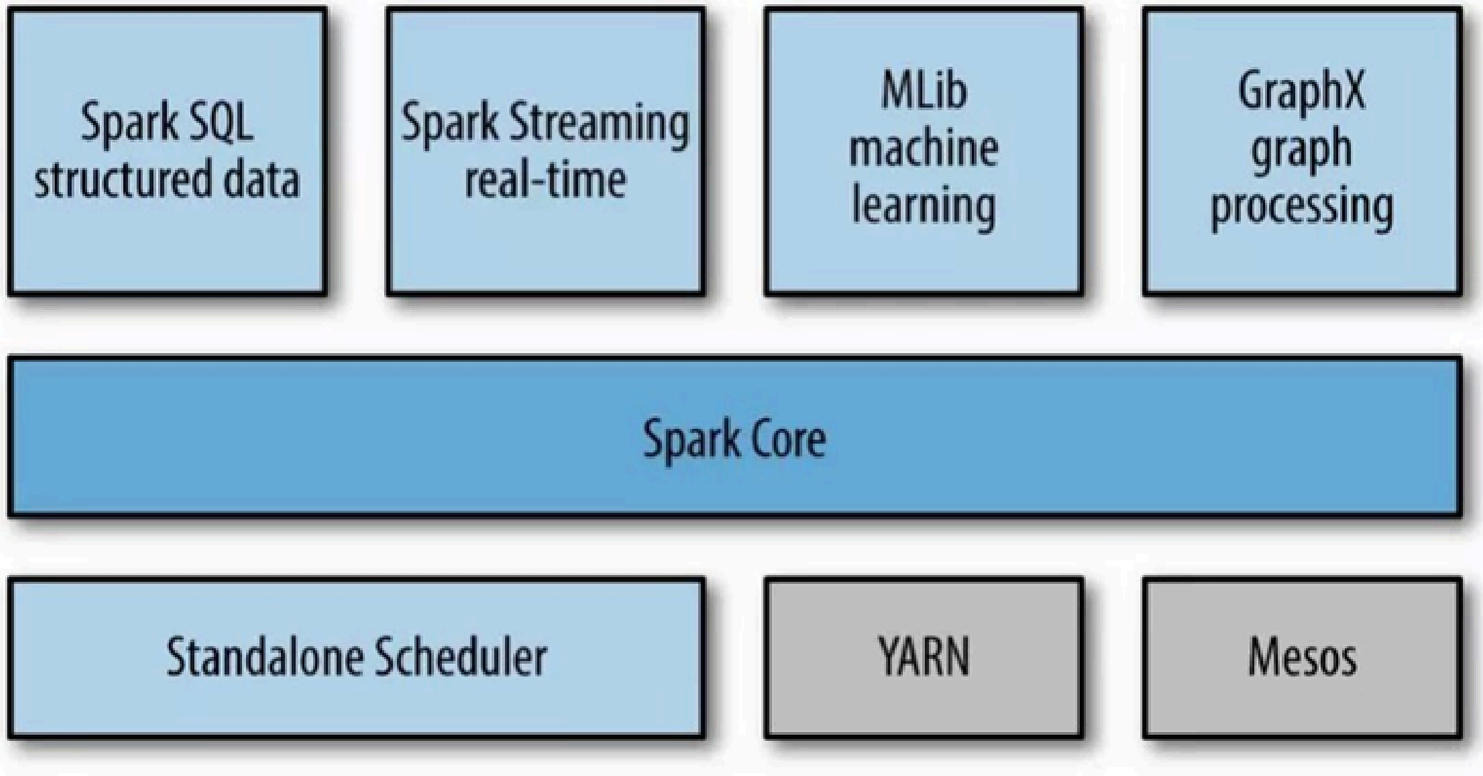
- Là một nền tảng xử lý dữ liệu phân tán mạnh mẽ, được thiết kế để hỗ trợ các tác vụ xử lý dữ liệu lớn nhanh chóng và hiệu quả
- Tính năng nổi bật: in-memory computing

1.4 KIẾN TRÚC APACHE SPARK



- Driver Program
- Cluster Manager
- Executor

1.4 CÁC THÀNH PHẦN CỦA APACHE SPARK



2.1 GIỚI THIỆU THUẬT TOÁN

- Thuật toán được thiết kế để giải quyết bài toán Clustering
- Thuật toán cơ bản được thiết kế để làm việc với dữ liệu dạng số

2.2 Ý TƯỞNG

- Nhóm các điểm dữ liệu có khoảng cách bé nhất với lại với nhau
- Khoảng cách thường được sử dụng là khoảng cách Euclidean

2.3 TRIỂN KHAI THUẬT TOÁN

- Input: dữ liệu X gồm n điểm dữ liệu và số lượng cụm cần tìm K
- Output: các điểm trung tâm M và vector nhãn cho từng điểm dữ liệu Y

2.3 TRIỂN KHAI THUẬT TOÁN

1. Chọn K điểm bất kỳ làm các trung tâm cụm ban đầu
2. Phân mỗi điểm dữ liệu vào cụm có trung tâm gần nó nhất
3. Nếu việc gán từng điểm dữ liệu vào cụm ở bước 2 không thay đổi so với vòng lặp trước thì ta dừng thuật toán
4. Cập nhật điểm trung tâm cho từng cụm bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cụm đó sau bước 2
5. Quay lại bước 2

3.1 DATASET

- Các giao dịch mua hàng được thực hiện cho một công ty bán lẻ trực tuyến có trụ sở tại Vương quốc Anh trong khoảng thời gian tám tháng

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

3.2 BÀI TOÁN

- Gợi ý các mặt hàng dựa trên nhu cầu mua hàng của người dùng sử dụng mô hình RFM (Recency, Frequency, Monetary)

3.3. THỰC HIỆN THUẬT TOÁN

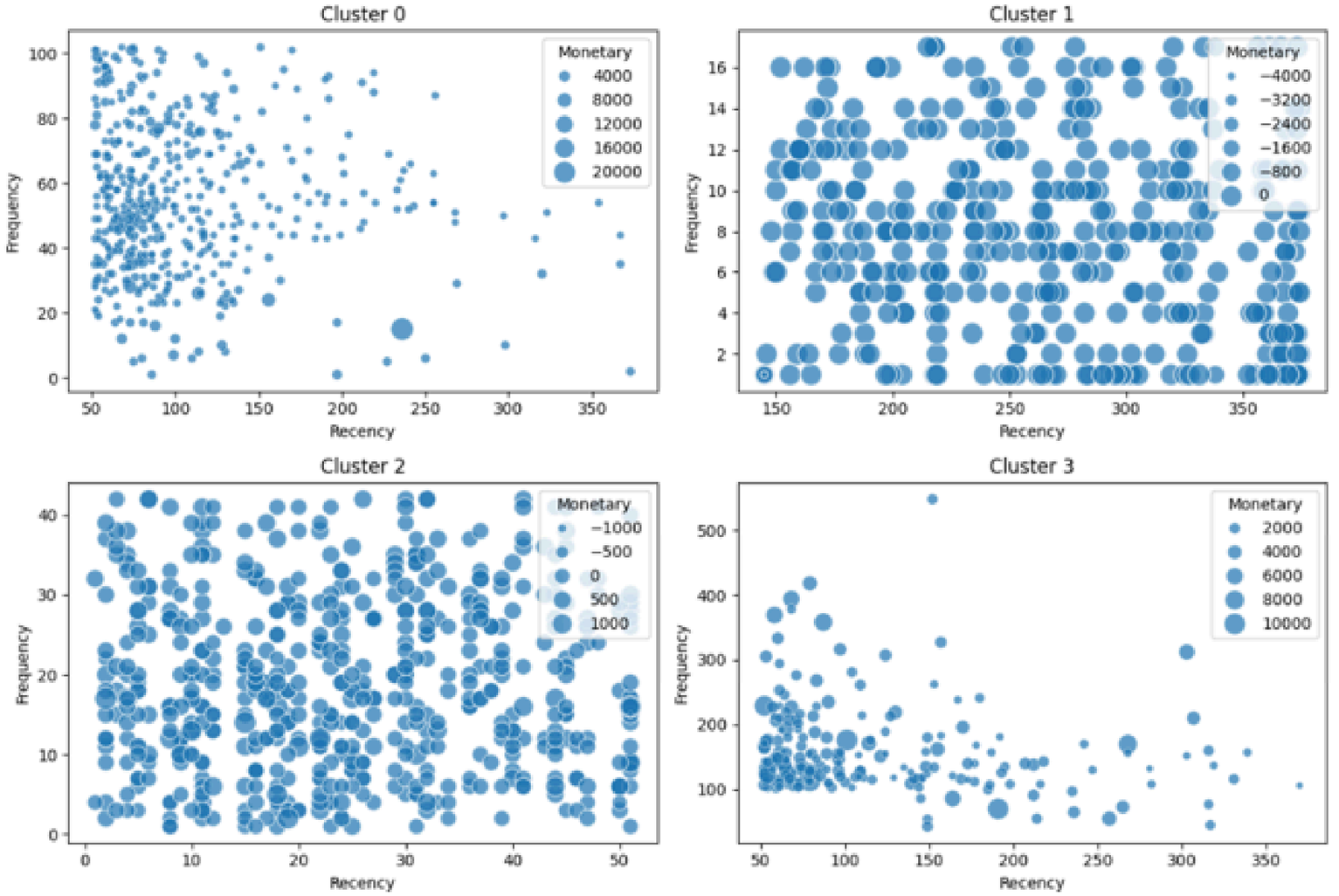
```
[27] kmeans = KMeans(k=10, maxIter=300, seed=42, featuresCol="scaledFeatures", predictionCol="prediction")
     model = kmeans.fit(scaled_df)
```

```
[41] predictions = model.transform(scaled_df)
     predictions = predictions.withColumnRenamed("prediction", "kmeans_cluster")
     predictions.show()
     result_df = predictions.select(*columns_for_clustering, "kmeans_cluster")
     result_df.show()
     centers = model.clusterCenters()
     print("Cluster Centers: ")
     for center in centers:
         print(center)
```

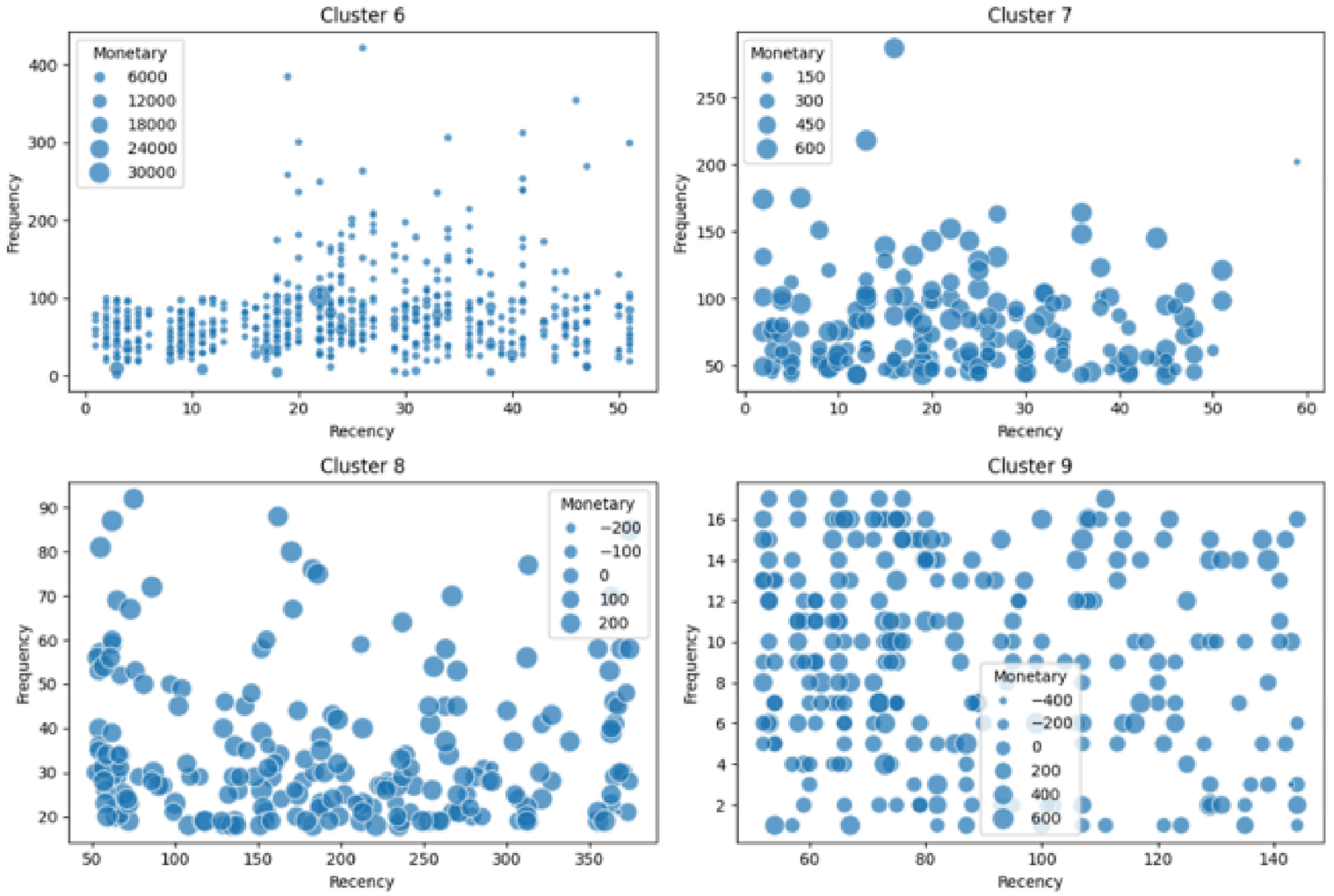
3.3. THỰC HIỆN THUẬT TOÁN

```
num_clusters = 10

cluster_recommendations = {}

for cluster_id in range(num_clusters):
    customers_in_cluster = predictions.filter(predictions.kmeans_cluster == cluster_id).select("CustomerID").distinct()

    cluster_transactions = dff.join(customers_in_cluster, on="CustomerID", how="inner")

    top_products_for_cluster = (
        cluster_transactions.groupBy("StockCode")
        .agg(F.count("InvoiceNo").alias("transaction_count"))
        .orderBy(F.desc("transaction_count"))
        .limit(10)
        .select("StockCode")
        .rdd.flatMap(lambda x: x)
        .collect()
    )

    cluster_recommendations[f"Cluster {cluster_id}"] = top_products_for_cluster

for cluster, recommended_products in cluster_recommendations.items():
    print(f"{cluster} -> Recommended Products: {recommended_products}")
```

3.4. KẾT QUẢ


Cluster 0
Frequency
Recency
Monetary
4000
8000
12000
16000
20000
Cluster 1
Frequency
Recency
Monetary
−4000
−3200
−2400
−1600
−800
0
Cluster 2
Frequency
Recency
Monetary
−1000
−500
0
500
1000
Cluster 3
Frequency
Recency
Monetary
2000
4000
6000
8000
10000

3.4. KẾT QUẢ


Cluster 6
Monetary
6000
12000
18000
24000
30000
Frequency
Recency
Cluster 7
Monetary
150
300
450
600
Frequency
Recency
Cluster 8
Monetary
−200
−100
0
100
200
Frequency
Recency
Cluster 9
Monetary
−400
−200
0
200
400
600
Frequency
Recency

3.4. KẾT QUẢ

```
Cluster 0 -> Recommended Products: ['22423', '85123A', '84879', '47566', 'POST', '22720', '22960', '85099B', '21212', '20725']
Cluster 1 -> Recommended Products: ['85123A', '22423', 'POST', '47566', '22457', '22427', '22969', '84879', '84946', '21232']
Cluster 2 -> Recommended Products: ['84879', 'POST', '22086', '22423', '22138', '23084', '23355', '85123A', '21034', '84946']
Cluster 3 -> Recommended Products: ['85123A', '22423', '47566', '22720', '84879', '85099B', '82482', '22960', '22961', '23298']
Cluster 4 -> Recommended Products: ['85123A', '85099B', '22423', '20725', '23203', '47566', '20727', '22197', '84879', '23209']
Cluster 5 -> Recommended Products: ['85123A', '22423', '47566', 'POST', '22720', '22960', '84879', '22457', '22470', '22178']
Cluster 6 -> Recommended Products: ['22423', '85123A', 'POST', '84879', '85099B', '23084', '22086', '22138', '20727', '47566']
Cluster 7 -> Recommended Products: ['21034', '22086', '23084', '85123A', '23321', '22197', '22469', '22578', '23301', '22910']
Cluster 8 -> Recommended Products: ['21034', '85123A', '47566', '22960', '22138', 'M', '22469', '22804', '22197', '22139']
Cluster 9 -> Recommended Products: ['84946', '84879', '85099B', '22423', '85123A', '47566', '22178', '22487', 'M', '21790']
```

4.1. KẾT LUẬN

Apache Spark không chỉ đơn thuần là một công cụ xử lý dữ liệu lớn, mà còn là một nền tảng mạnh mẽ hỗ trợ các tác vụ trong lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo

4.2 HƯỚNG PHÁT TRIỂN

- Thử nghiệm với những bộ dữ liệu có khối lượng lớn hơn
- Mở rộng hệ thống, tích hợp thêm CSDL NoSQL để tạo thành pipeline hoàn chỉnh

Thank you for listening