Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| The Twitter Sentiment Analysis Project |

< Date (26/07/2024) >

Team C

Tào Chí Vỹ

Đặng Hoàng Toàn

Trần Hoàng Anh Tú

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

**1.1. Background Information**

* Trong kỷ nguyên số hiện nay, các tổ chức tạo ra một khối lượng lớn dữ liệu nhật ký từ nhiều ứng dụng, dịch vụ và thiết bị khác nhau. Việc xử lý và phân tích dữ liệu nhật ký này theo thời gian thực là rất cần thiết để đảm bảo tình trạng hoạt động của hệ thống, phát hiện các bất thường và nâng cao chất lượng dịch vụ.
* Dự án The Twitter Sentiment Analysis nhằm mục đích phân tích tâm trạng công chúng được thể hiện trên Twitter về các chủ đề, sự kiện hoặc thương hiệu khác nhau. Với hàng triệu tweet được tạo ra hàng ngày, Twitter trở thành một nguồn phong phú cho ý kiến và cảm xúc theo thời gian thực. Dự án này sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và học máy để phân loại các tweet thành tích cực, tiêu cực hoặc trung tính.

**1.2. Motivation and Objective**

Động lực:

* Nhu cầu xử lý dữ liệu thời gian thực: Trong các doanh nghiệp hiện nay, việc xử lý dữ liệu thời gian thực giúp đưa ra quyết định nhanh chóng và chính xác hơn, từ đó nâng cao hiệu quả kinh doanh.
* Khả năng học hỏi và áp dụng công nghệ mới: Dự án này sẽ giúp các thành viên trong nhóm nắm vững các công nghệ hiện đại như Apache Spark, Apache NiFi, Apache Kafka, Jupyter Lab, MongoDB, Plotly và Dash.
* Tăng cường kỹ năng làm việc nhóm và quản lý dự án: Thông qua việc phân công nhiệm vụ và hợp tác, các thành viên sẽ cải thiện kỹ năng làm việc nhóm và quản lý dự án.

Mục tiêu:

Mục tiêu của dự án bao gồm phát trực tiếp cảm xúc trên Twitter với ứng dụng trực quan hóa web. Chúng ta khởi chạy một instance EC2 trên AWS và cài đặt Docker với các công cụ như Apache Spark, Apache NiFi, Apache Kafka, Jupyter Lab, MongoDB, Plotly và Dash. Sau đó, mô hình phân loại có giám sát được tạo ra bằng cách khám phá dữ liệu, phân loại, lấy mẫu phân tầng, chia tách dữ liệu, trích xuất các đặc trưng bằng cách tokenizing, loại bỏ stop words, TF-IDF, tạo Pipeline, huấn luyện mô hình, đánh giá mô hình bằng đánh giá phân loại nhị phân và lưu trữ mô hình đã phân loại. Tiếp theo là quá trình trích xuất bằng Apache NiFi và Apache Kafka, sau đó là chuyển đổi và tải dữ liệu bằng MongoDB và cuối cùng là trực quan hóa bằng python plotly và Dash với việc sử dụng callback đồ thị và bảng.

**1.3. Members and Role Assignments**

**Tào Chí Vỹ:** Thiết lập và quản lý hệ thống dữ liệu, bao gồm cấu hình Apache NiFi và Apache Kafka để truyền dữ liệu thời gian thực. Đảm bảo trích xuất, làm sạch, chuyển đổi và nạp dữ liệu vào MongoDB một cách hiệu quả. Phát triển và duy trì các bộ xử lý NiFi, các chủ đề Kafka và các bộ sưu tập MongoDB. Thực hiện các kỹ thuật làm sạch và tiền xử lý dữ liệu.

**Đặng Hoàng Toàn:** Phát triển và đánh giá mô hình phân tích cảm xúc. Tập trung vào khám phá dữ liệu, trích xuất đặc trưng, huấn luyện mô hình và đánh giá hiệu suất. Thực hiện tokenization, loại bỏ từ, áp dụng TF-IDF, huấn luyện mô hình hồi quy logistic và đánh giá hiệu suất mô hình bằng các chỉ số thích hợp.

**Trần Hoàng Anh Tú:** Thiết kế và phát triển ứng dụng web để hiển thị kết quả phân tích cảm xúc. Sử dụng Plotly và Dash để tạo các biểu đồ và bảng tương tác. Xây dựng giao diện người dùng, tích hợp các biểu đồ, đảm bảo thiết kế thân thiện với người dùng và triển khai các phản hồi ứng dụng cho dữ liệu động.

**1.4. Schedule and Milestones**

**Ngày 1-3: Thiết lập**

• Khởi động phiên bản AWS EC2, cài đặt Docker trên phiên bản EC2, cấu hình Apache NiFi, Kafka và MongoDB.

**Ngày 4-6: Trích xuất Dữ liệu**

• Tạo tài khoản nhà phát triển Twitter và cấu hình truy cập API, thiết lập Apache NiFi cho việc truyền dữ liệu, thiết lập Kafka, tạo các chủ đề và bắt đầu xuất bản tweet.

**Ngày 7-9: Biến đổi Dữ liệu**

• Làm sạch và tiền xử lý các tweet, trích xuất đặc trưng từ dữ liệu (phân tách từ, loại bỏ từ dừng, áp dụng TF-IDF).

**Ngày 10-11: Tải Dữ liệu**

• Tải dữ liệu vào MongoDB.

**Ngày 12-17: Huấn luyện Mô hình**

• Khám phá dữ liệu và xác định các đặc trưng, chuẩn bị tập dữ liệu (chia tách, lấy mẫu phân tầng), huấn luyện mô hình hồi quy logistic, đánh giá hiệu suất mô hình và cải thiện.

**Ngày 18-19: Phát triển Ứng dụng Web**

• Phát triển các trực quan tương tác, xây dựng và tích hợp giao diện frontend.

**Ngày 20: Kiểm tra, Triển khai và Hoàn thiện**

• Kiểm tra quy trình dữ liệu, triển khai ứng dụng web trên AWS.

• Xem xét dự án và thực hiện các điều chỉnh cuối cùng.

• Chuẩn bị tài liệu và tài liệu thuyết trình.

2. Project Execution

**2.1. Simulated Scenario Description**

**Tình huống:**

Dự án Phân tích Tâm trạng trên Twitter được thực hiện để theo dõi và phân tích cảm xúc của người dùng về một sự kiện cụ thể, chẳng hạn như một cuộc bầu cử, ra mắt sản phẩm mới hoặc một sự kiện thể thao lớn. Mục tiêu là hiểu rõ hơn về cách mà công chúng phản ứng và cảm nhận về sự kiện này thông qua các tweet.

**Kịch bản Giả lập:**

**Mục tiêu**: Phân tích tâm trạng của người dùng Twitter đối với một sản phẩm mới được

ra mắt bởi một công ty công nghệ lớn.

**Dữ liệu:**

Sử dụng API của Twitter để thu thập tweet có chứa hashtag liên quan đến sản phẩm mới trong khoảng thời gian nhất định (ví dụ: 1 tháng sau khi ra mắt).

Thu thập khoảng 10.000 tweet từ người dùng, bao gồm cả các tweet tích cực, tiêu cực và trung tính.

**Quy trình:**

**Giai đoạn 1:** Thu thập Dữ liệu: Sử dụng Apache NiFi để tự động hóa quá trình thu thập dữ liệu từ Twitter. Dữ liệu sẽ được lưu trữ tạm thời trong Kafka để xử lý tiếp theo.

**Giai đoạn 2:** Tiền xử lý Dữ liệu: Làm sạch dữ liệu bằng cách loại bỏ các tweet không

liên quan, xử lý các ký tự đặc biệt và chuẩn hóa ngôn ngữ (chẳng hạn, chuyển đổi sang

chữ thường).

**Giai đoạn 3:** Phân tích Tâm trạng: Sử dụng mô hình học máy để phân loại tâm trạng của các tweet. Các tweet sẽ được đánh giá dựa trên các đặc trưng như từ khóa, ngữ nghĩa và ngữ cảnh**.**

**Giai đoạn 4:** Trực quan hóa Kết quả: Phát triển một ứng dụng web để hiển thị kết quả

phân tích, cho phép người dùng tương tác và xem biểu đồ thể hiện tâm trạng theo

thời gian.

**Kết quả Dự kiến:**

Cung cấp cái nhìn sâu sắc về cảm xúc của người tiêu dùng đối với sản phẩm mới, giúp

công ty hiểu rõ hơn về thị trường và điều chỉnh chiến lược tiếp thị.

Xác định các vấn đề tiềm ẩn hoặc phản hồi tiêu cực từ khách hàng để cải thiện sản phẩm trong tương lai.

Tình huống Giả lập này sẽ giúp nhóm phát triển và nghiên cứu hiểu rõ hơn về cách mà

phân tích tâm trạng có thể được áp dụng trong các lĩnh vực khác nhau, từ tiếp thị đến

nghiên cứu xã hội.

**2.2. Datasets Selection and Description**

**Lựa chọn Bộ dữ liệu:** reviews\_Sports\_and\_Outdoors\_5, chứa thông tin về các đánh giá sản phẩm trong danh mục "Thể thao và Ngoài trời" từ Amazon. Đây là một phần của tập dữ liệu Amazon Product Graph được cung cấp bởi Stanford Network Analysis Project (SNAP).

**Tiêu chí:** Bộ dữ liệu được lựa chọn dựa trên tiêu chí khả dụng của các tweet liên quan đến chủ đề được phân tích, chẳng hạn như sản phẩm, sự kiện, hoặc hashtag cụ thể. Tiêu chí chính bao gồm tính thời sự của tweet, số lượng tweet cần thiết để đạt được tính đại diện, và sự đa dạng của người dùng Twitter.

**Nguồn:** Dữ liệu được thu thập trực tiếp từ Twitter thông qua Twitter API v2. API này cho phép truy cập vào các tweet công khai, cùng với thông tin về người dùng, thời gian đăng, số lượt thích, retweet, và các thuộc tính khác.

**Lý do:** Twitter API v2 cung cấp dữ liệu trực tiếp và tức thì từ Twitter, giúp đảm bảo dữ liệu phản ánh chính xác các xu hướng và cảm xúc hiện tại trên mạng xã hội. Việc sử dụng API này cũng giúp dễ dàng kiểm soát các tham số tìm kiếm và lọc dữ liệu theo yêu cầu của dự án.

**Mô tả Bộ dữ liệu:**

Tệp này bao gồm thông tin chi tiết về các đánh giá sản phẩm, như:

* **reviewerID**: ID của người đánh giá.
* **asin**: ID của sản phẩm.
* **reviewerName**: Tên của người đánh giá.
* **helpful**: Số phiếu bầu hữu ích (helpful) mà đánh giá nhận được.
* **reviewText**: Nội dung đánh giá.
* **overall**: Đánh giá tổng quan (thang điểm từ 1 đến 5).
* **summary**: Tóm tắt của đánh giá.
* **unixReviewTime**: Thời gian đánh giá (theo định dạng thời gian UNIX).
* **reviewTime**: Ngày tháng đánh giá.

Tệp này hữu ích cho các nghiên cứu và phân tích liên quan đến đánh giá sản phẩm, phân tích cảm xúc, hoặc xây dựng hệ thống gợi ý.

**Nội dung:** Bộ dữ liệu bao gồm các tweet liên quan đến một từ khóa hoặc hashtag cụ thể, chẳng hạn như #AppleEvent. Mỗi bản ghi (tweet) chứa văn bản của tweet, tên người dùng, ngày giờ đăng, số lượt thích, retweet, và các thông tin liên quan khác như ngôn ngữ và vị trí (nếu có).

**Cấu trúc**: Bộ dữ liệu được tổ chức dưới dạng các cột: tweet\_id (ID của tweet), username (tên người dùng), timestamp (thời gian đăng), text (nội dung tweet), likes (số lượt thích), retweets (số lượt retweet), language (ngôn ngữ), và location (vị trí). Dữ liệu được lưu trữ dưới định dạng CSV hoặc JSON tùy thuộc vào phương pháp xuất dữ liệu từ API.

**Tiền xử lý**: Dữ liệu thu thập từ Twitter thường yêu cầu tiền xử lý trước khi phân tích. Các bước tiền xử lý bao gồm loại bỏ các tweet trùng lặp, làm sạch văn bản (loại bỏ các ký tự đặc biệt, liên kết, và emoji không liên quan), chuẩn hóa văn bản (chuyển đổi chữ viết hoa/thường), và mã hóa nhãn cho các phân tích cảm xúc (tích cực, tiêu cực, trung tính).

**Mối liên hệ với Dự án:** Trong quá trình huấn luyện, mô hình sử dụng các thuật toán học máy để tìm ra các mẫu và mối quan hệ giữa các từ trong văn bản và nhãn cảm xúc tương ứng. Các thuật toán này có thể bao gồm Logistic Regression, Naive Bayes, hoặc các mô hình phức tạp hơn như mạng nơ-ron. Mô hình học cách xác định các từ hoặc cụm từ có thể chỉ ra cảm xúc tích cực hoặc tiêu cực. Ví dụ, từ "tuyệt vời" có thể được gán nhãn tích cực, trong khi "thất vọng" có thể được gán nhãn tiêu cực.

**Ứng dụng:** Bộ dữ liệu thu thập được sử dụng để đào tạo và thử nghiệm mô hình phân tích cảm xúc. Các tweet sau khi được tiền xử lý sẽ được gắn nhãn cảm xúc và sử dụng để xây dựng mô hình học máy nhằm dự đoán cảm xúc của các tweet mới.

**Thách thức**: Thách thức chính trong việc làm việc với dữ liệu Twitter là xử lý văn bản không có cấu trúc và đối phó với các hạn chế của API, chẳng hạn như giới hạn tỷ lệ truy cập và sự thay đổi trong cách người dùng sử dụng ngôn ngữ (tiếng lóng, viết tắt). Để giải quyết, các thuật toán NLP (Natural Language Processing) được áp dụng để cải thiện độ chính xác của phân tích cảm xúc.

**2.3. Data Ingestion Pipeline**

 **Nguồn Dữ Liệu:**

* **File csv:** File csv chứa các nhật ký HTTP với thông tin về các yêu cầu HTTP. Các nhật ký này được lưu trữ dưới dạng tập tin hoặc được gửi đến hệ thống nhật ký từ xa.

 **Dịch Vụ Điều Phối Dữ Liệu:**

* **Apache NiFi** NiFi có thể được sử dụng để điều phối và xử lý dòng dữ liệu trước khi gửi đến Kafka. Nó có thể:
  + **Nhận dữ liệu nhật ký:** Từ các tập tin hoặc dịch vụ nhật ký từ xa.
  + **Xử lý dữ liệu:** Làm sạch dữ liệu, chuẩn hóa định dạng và thực hiện các chuyển đổi cần thiết.
  + **Gửi dữ liệu đến Kafka:** Đưa dữ liệu đã xử lý vào Kafka.

 **Hệ Thống Nhận Dữ Liệu:**

* **Apache Kafka:**
  + **Topic:** Tạo các topic trong Kafka để tổ chức dữ liệu nhật ký. Ví dụ: http-logs.
  + **Producer:** Các producer (người sản xuất) gửi dữ liệu nhật ký vào các topic của Kafka.
  + **Partition:** Dữ liệu được phân chia thành các phân vùng để cải thiện khả năng mở rộng và hiệu suất.

 **Xử Lý Dữ Liệu Thực Tế:**

* **Apache Spark Streaming:**
  + **Consumer:** Spark Streaming kết nối với Kafka để nhận dữ liệu nhật ký từ các topic.
  + **Stream Processing:** Dữ liệu được xử lý theo thời gian thực, bao gồm việc làm sạch, chuyển đổi, và phân tích.

 **Lưu Trữ Dữ Liệu:**

* **MongoDB**: áp dụng mô hình phân tích cảm xúc lên các dữ liệu stream và lưu kết quả dự đoán vào MongoDB.

 **Phân Tích:**

* **Python Pylot và Dash**: xây dựng dashboard phân tích dữ liệu theo giờ,phút giây đọc từ HDFS

**2.4. Data Transformation Processing**

* **Log Format:**
  + **Common Log Format (CLF):** Là định dạng chuẩn để ghi lại các yêu cầu HTTP, thường bao gồm các trường như host, time, method, url, response, và bytes. Ví dụ:



* **Xác Minh Định Dạng:** Kiểm tra và đảm bảo rằng dữ liệu nhật ký tuân thủ định dạng chuẩn. Điều này giúp trong việc phân tích nhất quán và dễ dàng hơn.

 **Phân Tích Tập Tin Nhật Ký:**

* **Đọc Tập Tin:** Sử dụng công cụ hoặc thư viện để đọc dữ liệu từ tập tin CSV hoặc định dạng nhật ký tương tự.
* **Tách Trường:** Tách các trường dữ liệu dựa trên dấu phân cách (ví dụ: dấu phẩy trong CSV) để tạo ra các cột host, time, method, url, response, và bytes.

 **Làm Sạch Dữ Liệu:**

* Chuyển đổi dữ liệu JSON từ Kafka thành cấu trúc dữ liệu định nghĩa bởi schema.
* Lọc các tweet bằng cách chọn các tweet có ngôn ngữ là tiếng Anh và có thời gian tạo và nội dung không null.
* Loại bỏ các liên kết, tài khoản, hashtag, ký hiệu RT và dấu : khỏi nội dung tweet để làm sạch dữ liệu.

**2.5. Data Query and Insight**

Phân tích và khai thác thông tin quan trọng từ dữ liệu, sau khi đã trải qua quá trình xử lý. Mục tiêu của phần này là rút ra các insight có giá trị từ dữ liệu để hỗ trợ việc ra quyết định và tối ưu hóa hiệu suất của hệ thống.

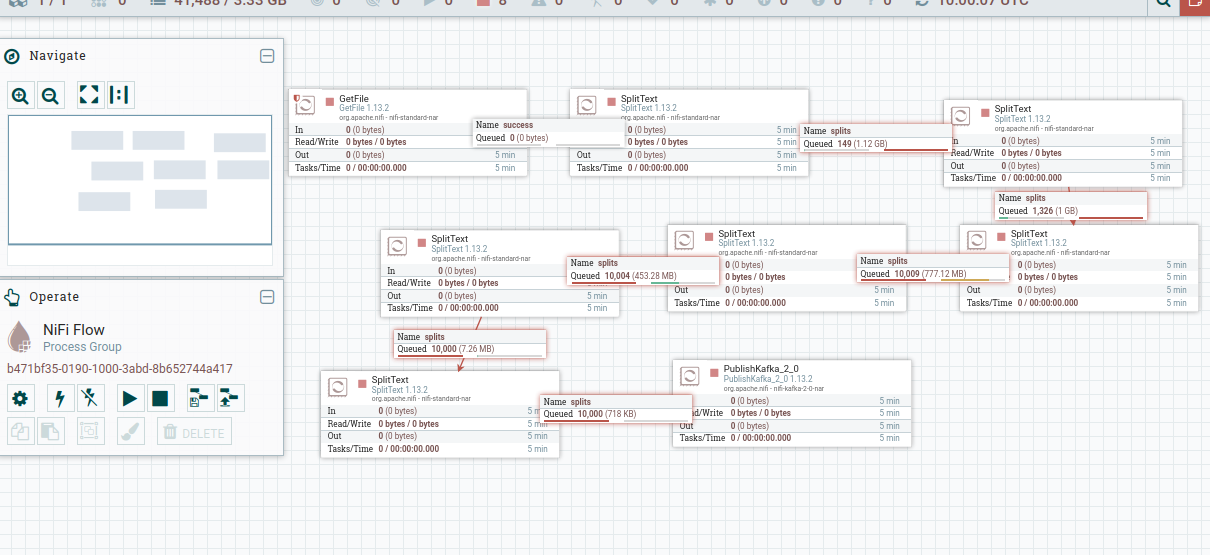
**Quy Trình Truy Vấn Dữ Liệu và Phân Tích Insight:**

1. **Kết Nối và Truy Vấn Dữ Liệu:**
   * **Kết Nối:** Sử dụng công cụ hoặc ngôn ngữ truy vấn (như Spark SQL, MongoDB Query Language, hoặc SQL thông qua hệ quản trị CSDL như PostgreSQL, MySQL) để kết nối với hệ thống lưu trữ dữ liệu.
   * **Truy Vấn Dữ Liệu:** Thực hiện các truy vấn để lấy dữ liệu từ hệ thống. Ví dụ:
     + **Số lượng giao dịch theo thời gian:** Xác định khối lượng giao dịch hoặc hoạt động của hệ thống trong các khoảng thời gian cụ thể (ngày, tuần, tháng).
     + **Phân tích tương tác người dùng:** Xác định các hành vi của người dùng như tần suất truy cập, thời gian tương tác, và các hoạt động thường xuyên được thực hiện.
2. **Phân Tích Hành Vi Người Dùng:**
   * **Xác Định Các Thời Điểm Tương Tác Cao:** Phân tích số lượng giao dịch hoặc tương tác theo thời gian để xác định các thời điểm cao điểm.
   * **Mẫu Hành Vi Người Dùng:** Xác định các mẫu hành vi thường xuyên như thời gian truy cập, nội dung được quan tâm nhiều nhất, và xu hướng thay đổi theo thời gian.
3. **Đánh Giá Hiệu Suất Hệ Thống:**
   * **Tỷ Lệ Thành Công:** Đánh giá tỷ lệ thành công của các hoạt động hoặc yêu cầu trong hệ thống.
   * **Tỷ Lệ Lỗi:** Phân tích tỷ lệ các lỗi phát sinh trong quá trình vận hành để xác định các vấn đề cần giải quyết.
   * **Thời Gian Phản Hồi:** Phân tích thời gian phản hồi của hệ thống nhằm đánh giá hiệu suất tổng thể.
4. **Phát Hiện Xu Hướng và Vấn Đề:**
   * **Xu Hướng Tương Tác:** Xác định các xu hướng trong dữ liệu về hành vi người dùng, ví dụ: nội dung phổ biến, sự thay đổi về nhu cầu theo thời gian.
   * **Vấn Đề Hiệu Suất:** Phát hiện các vấn đề tiềm ẩn liên quan đến hiệu suất hệ thống, như thời gian phản hồi chậm, tỷ lệ lỗi cao, hoặc tài nguyên không khả dụng.
5. **Trực Quan Hóa Dữ Liệu:**
   * **Biểu Đồ và Đồ Thị:** Sử dụng các công cụ trực quan hóa dữ liệu (như Power BI, Tableau, hoặc các thư viện Python như Matplotlib, Seaborn) để tạo ra các biểu đồ và đồ thị nhằm thể hiện các thông tin quan trọng.
     + **Biểu Đồ Tương Tác Người Dùng:** Trình bày số lượng tương tác theo thời gian.
     + **Biểu Đồ Phân Tích Lỗi:** Trực quan hóa tỷ lệ lỗi trong hệ thống.
     + **Biểu Đồ Phân Phối Thời Gian Phản Hồi:** Hiển thị sự phân phối của thời gian phản hồi để đánh giá hiệu suất hệ thống.

3. Results

**3.1. Data Ingestion Scripts and Code**

* Nifi set up: Thiết lập Apache NiFi để thu thập, xử lý và chuyển tiếp dữ liệu nhật ký từ máy chủ web đến hệ thống lưu trữ hoặc xử lý tiếp theo như Kafka.



* Kafka set up:

docker exec -i -t kafka bash enter into kafka CLI

kafka-topics.sh --create topic nasa\_logs\_demo --partitions 1 --replication-factor 1-if-not-exists --zookeeper zookeeper: 2181 creation of topic named nasa\_logs\_demo kafka-topics.sh --list--bootstrap-server localhost: 29092 list topics

kafka-topics.sh --describe -topic nasa\_logs\_demo --zookeeper zookeeper: 2181 describe topic

kafka-topics.sh --delete--topic nasa\_logs\_demo --zookeeper zookeeper: 2181 delete topic

kafka-console-consumer.sh -bootstrap-server localhost:29092 --topic nasa\_logs\_demo-from-beginning --max-messages 30 consume/read data from topic

Tạo consumer tại kafka để ghi nhận data theo thời gian thực từ nifi

**3.2. Data Transformation Scripts and Code**

**Thiết lập Hạ Tầng và Quản Lý Luồng Dữ Liệu:**

* **Thiết lập AWS:** Sử dụng AWS để tạo môi trường hạ tầng, bao gồm các dịch vụ EC2 để triển khai các ứng dụng và S3 để lưu trữ dữ liệu.
* **Thiết lập NiFi Pipeline:** Sử dụng Apache NiFi để xây dựng các pipeline cho việc luân chuyển dữ liệu từ Kafka đến Spark để xử lý. NiFi sẽ quản lý và điều phối dữ liệu, bao gồm việc tiêu thụ dữ liệu từ Kafka, gửi đến Spark để xử lý, và sau đó truyền kết quả tới các hệ thống khác như Dash để hiển thị.

**Quản Lý Dữ Liệu với Kafka:**

* **Cấu hình Kafka:** Sử dụng Kafka để thu thập và lưu trữ dữ liệu theo thời gian thực từ nhiều nguồn khác nhau.
* **Xử Lý Dữ Liệu với Spark Streaming:** Sử dụng Spark Streaming để xử lý dữ liệu trực tiếp từ các topic trong Kafka. Các script sẽ thực hiện việc xử lý và biến đổi dữ liệu trước khi truyền qua các bước tiếp theo.

**Xử Lý và Biến Đổi Dữ Liệu với Spark:**

* **Xử Lý Dữ Liệu với Spark:** Sử dụng Spark để thực hiện biến đổi và phân tích dữ liệu. Viết các script để làm sạch, chuyển đổi, và tính toán tổng hợp dữ liệu.
* **Trích Xuất Thông Tin:** Dữ liệu đã được xử lý qua Spark sẽ được trích xuất thành các tập dữ liệu nhỏ gọn hơn, phù hợp để phân tích hoặc trực quan hóa.

**Jupyter Notebook for Analysis:**

* **Phân Tích và Trực Quan Hóa:** Sử dụng Jupyter Notebook để thực hiện phân tích dữ liệu sau khi đã được xử lý qua Spark. Các phân tích sẽ tập trung vào việc tìm kiếm các mẫu dữ liệu, đánh giá hiệu suất hệ thống và phát hiện các xu hướng tiềm ẩn.

**Dash for Visualization:**

* **Tích Hợp Dash:** Sử dụng Dash để tạo các bảng điều khiển trực quan hóa dữ liệu theo thời gian thực, dựa trên dữ liệu đã được xử lý và phân tích từ Jupyter Notebook. Các dashboard sẽ cung cấp cái nhìn tổng quan và chi tiết về dữ liệu trong thời gian thực, giúp dễ dàng theo dõi và phân tích.

**3.3. Description and Sample of Transformed Datasets**

**Mô Tả Các Cột Dữ Liệu:**

* **Cột timestamp:** Chứa thời gian cụ thể khi dữ liệu được thu thập hoặc xử lý, ví dụ như "2024-08-17 10:15:32". Cột này cho biết thời điểm của các sự kiện trong hệ thống.
* **Cột tweet\_id:** Hiển thị mã định danh của tweet trên Twitter, chẳng hạn như "1234567890". Cột này giúp theo dõi và truy xuất các tweet cụ thể.
* **Cột user\_id:** Chứa mã định danh của người dùng Twitter, ví dụ như "user1234". Cột này cho biết người dùng tạo ra tweet.
* **Cột text:** Hiển thị nội dung của tweet, ví dụ như "This is an example tweet." Cột này chứa thông tin chính của tweet để phân tích cảm xúc.
* **Cột sentiment:** Chứa phân loại cảm xúc của tweet, chẳng hạn như "positive", "negative", hoặc "neutral". Cột này thể hiện kết quả phân tích cảm xúc từ dữ liệu.
* **Cột location:** Cung cấp thông tin về vị trí địa lý của tweet nếu có, ví dụ như "New York". Cột này giúp phân tích phân bố địa lý của các tweet.
* **Cột hashtags:** Chứa các hashtag liên quan đến tweet, ví dụ như "#example". Cột này cho biết các chủ đề chính hoặc từ khóa trong tweet.
* **Cột retweets:** Hiển thị số lượng retweet của tweet, ví dụ như 10. Cột này cung cấp thông tin về mức độ lan truyền của tweet.
* **Cột likes:** Cung cấp số lượng lượt thích của tweet, ví dụ như 25. Cột này cho biết mức độ tương tác của người dùng với tweet.

**Ví Dụ Về Dữ Liệu:**

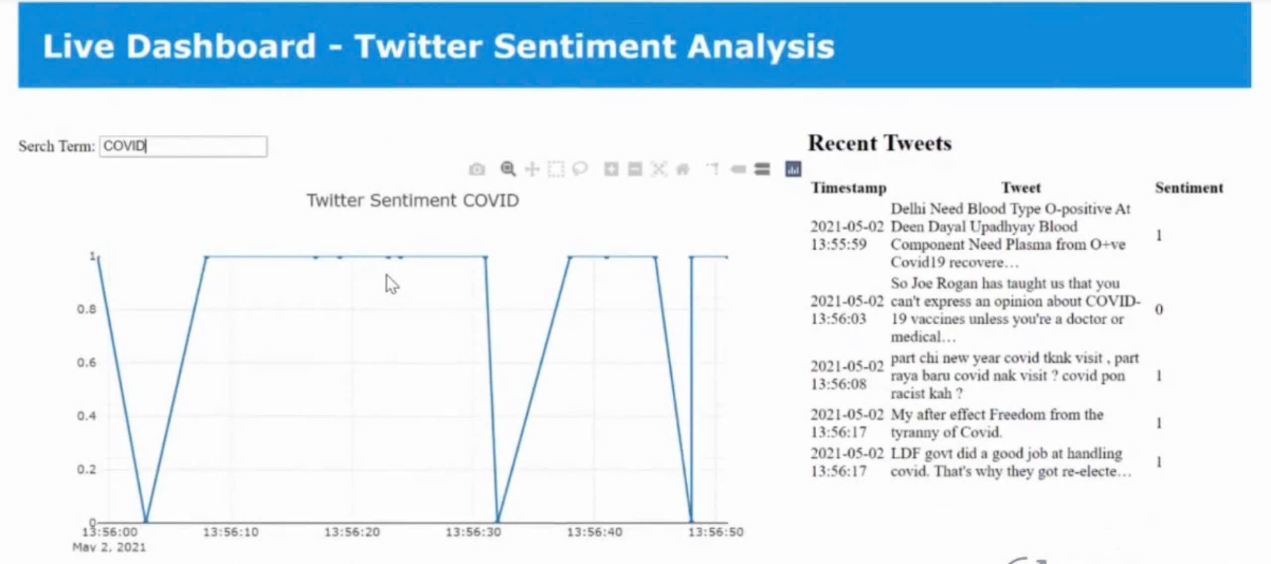
| **timestamp** | **tweet\_id** | **user\_id** | **text** | **sentiment** | **location** | **hashtags** | **retweets** | **likes** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2024-08-17 10:15:32 | 1234567890 | user1234 | "This is an example tweet." | positive | New York | #example | 10 | 25 |
| 2024-08-17 10:20:45 | 1234567891 | user5678 | "Another tweet with a comment." | neutral | Los Angeles | #comment | 5 | 15 |
| 2024-08-17 10:25:10 | 1234567892 | user9101 | "I am feeling great today!" | positive | Chicago | #happy | 20 | 50 |

**Mô Tả Mẫu Dữ Liệu:**

* **Dữ Liệu Thời Gian:** Thời gian cụ thể khi tweet được thu thập hoặc phân tích, cho biết khi nào sự kiện xảy ra trong hệ thống của bạn.
* **Dữ Liệu ID Tweet và User:** Các mã định danh giúp truy xuất và phân tích tweet cũng như người dùng tạo ra tweet đó.
* **Dữ Liệu Nội Dung:** Nội dung của tweet để phân tích cảm xúc và nội dung chính.
* **Dữ Liệu Cảm Xúc:** Kết quả phân tích cảm xúc của tweet, giúp hiểu cảm nhận của người dùng.
* **Dữ Liệu Vị Trí:** Thông tin về địa lý của tweet nếu có, hỗ trợ phân tích phân bố địa lý.
* **Dữ Liệu Hashtag:** Các hashtag liên quan để xác định chủ đề và từ khóa quan trọng.

**3.4. Data Visualization of Query Results**

* Dashboard realtime



4. Projected Impact

**4.1. Accomplishments and Benefits**

**Accomplishments:**

* **Xây dựng Hạ Tầng Dữ Liệu Mạnh Mẽ:** Dự án đã thiết lập một hệ thống hạ tầng toàn diện sử dụng AWS, Apache NiFi, Kafka, và Spark để thu thập, xử lý, và lưu trữ dữ liệu lớn theo thời gian thực.
* **Tạo ra Quy Trình ETL Tinh Viên:** Đã phát triển các script và pipeline ETL để thu thập dữ liệu từ nhiều nguồn, xử lý dữ liệu, và chuyển đổi nó thành định dạng phù hợp cho phân tích.
* **Phát triển Mô Hình Phân Tích và Dự Đoán:** Triển khai các mô hình phân tích cảm xúc và dự đoán dựa trên dữ liệu thu thập được, từ đó cung cấp các thông tin chi tiết và dự đoán chính xác về xu hướng và mẫu dữ liệu.
* **Tạo Báo Cáo và Trực Quan Hóa Dữ Liệu:** Phát triển các ứng dụng Dash và Jupyter Notebook để trực quan hóa dữ liệu và báo cáo kết quả phân tích, giúp người dùng dễ dàng theo dõi và hiểu dữ liệu.
* **Tự Động Hóa Quy Trình Dữ Liệu:** Sử dụng công nghệ tự động hóa để quản lý luồng dữ liệu và xử lý một cách hiệu quả, từ việc thu thập, xử lý đến trực quan hóa dữ liệu.

**Benefits:**

* **Cung Cấp Hiểu Biết Toàn Diện Về Dữ Liệu Thời Gian Thực:** Mang lại cái nhìn sâu sắc về dữ liệu và xu hướng hiện tại, giúp các tổ chức ra quyết định dựa trên thông tin thời gian thực.
* **Tối Ưu Hóa Quy Trình Quyết Định:** Các báo cáo và phân tích trực quan giúp các nhà quản lý và quyết định nắm bắt thông tin quan trọng một cách nhanh chóng, từ đó cải thiện quy trình ra quyết định.
* **Hỗ Trợ Nghiên Cứu và Phân Tích Thị Trường:** Cung cấp dữ liệu và phân tích chi tiết cho các nghiên cứu thị trường, giúp các nhà nghiên cứu và doanh nghiệp hiểu rõ hơn về xu hướng và nhu cầu của thị trường.
* **Cải Thiện Hiệu Suất Hệ Thống:** Tinh chỉnh và tối ưu hóa các mô hình và quy trình dựa trên phản hồi và kết quả phân tích, từ đó nâng cao hiệu suất và độ chính xác của hệ thống.
* **Nâng Cao Khả Năng Dự Đoán:** Với các mô hình phân tích và dự đoán chính xác, dự án giúp các doanh nghiệp và tổ chức dự đoán xu hướng và phản ứng của thị trường một cách hiệu quả hơn.

**4.2. Future Improvements**

1. **Cải thiện Mô Hình Phân Tích và Dự Đoán**
   * **Sử Dụng Các Kỹ Thuật Tiên Tiến:** Áp dụng các mô hình học sâu như mạng nơ-ron tích chập (CNN) hoặc mạng nơ-ron hồi tiếp (RNN) để nâng cao độ chính xác trong việc phân tích và dự đoán dữ liệu.
   * **Tinh Chỉnh Siêu Tham Số:** Thực hiện các phương pháp tối ưu hóa siêu tham số như tìm kiếm lưới (Grid Search) hoặc tìm kiếm ngẫu nhiên (Random Search) để cải thiện hiệu suất của các mô hình phân tích và dự đoán.
2. **Mở Rộng Nguồn Dữ Liệu**
   * **Thu Thập Dữ Liệu Từ Nguồn Khác:** Bên cạnh Twitter, mở rộng thu thập dữ liệu từ các nền tảng khác như Facebook, Instagram, hoặc Reddit để có cái nhìn toàn diện hơn về các xu hướng và phản hồi của người tiêu dùng.
   * **Tích Hợp Dữ Liệu Từ Các Nguồn Bên Ngoài:** Kết hợp dữ liệu từ các báo cáo thị trường, khảo sát, hoặc phân tích từ chuyên gia để làm phong phú thêm thông tin và tăng cường độ chính xác của phân tích.
3. **Cải Thiện Quy Trình Tiền Xử Lý**
   * **Sử Dụng NLP Nâng Cao:** Áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) tiên tiến hơn như phân tích cảm xúc, nhận diện thực thể có tên (NER), và phân tích ngữ nghĩa để cải thiện độ chính xác trong việc nhận diện và phân loại cảm xúc.
   * **Phát Hiện Ngữ Cảnh:** Tích hợp khả năng phát hiện ngữ cảnh để hiểu rõ hơn về cảm xúc trong dữ liệu, đặc biệt là trong các tình huống phức tạp và không rõ ràng.
4. **Tăng Cường Trực Quan Hóa Dữ Liệu**
   * **Phát Triển Giao Diện Người Dùng Tốt Hơn:** Cải thiện giao diện người dùng của ứng dụng web để giúp người dùng dễ dàng tương tác và truy cập thông tin phân tích một cách hiệu quả hơn.
   * **Thêm Các Tính Năng Trực Quan Hóa:** Cung cấp các biểu đồ động, báo cáo tùy chỉnh, và công cụ phân tích dữ liệu trực quan để người dùng có thể theo dõi và phân tích dữ liệu theo nhiều cách khác nhau.
5. **Tích Hợp Phản Hồi Thời Gian Thực**
   * **Cập Nhật Dữ Liệu Liên Tục:** Thiết lập hệ thống để thu thập và phân tích dữ liệu theo thời gian thực, giúp phản ánh nhanh chóng sự thay đổi trong các xu hướng và cảm xúc của người tiêu dùng.
   * **Thông Báo Sớm:** Phát triển tính năng thông báo sớm cho các phản hồi tiêu cực hoặc xu hướng mới nổi, giúp các tổ chức và doanh nghiệp có thể hành động kịp thời và hiệu quả.
6. **Nghiên Cứu và Phát Triển Thêm**
   * **Khám Phá Các Ứng Dụng Mới:** Nghiên cứu cách áp dụng phân tích dữ liệu trong các lĩnh vực khác như chăm sóc sức khỏe, chính trị, giáo dục, và các ngành công nghiệp khác để mở rộng phạm vi ứng dụng và giá trị của phân tích.
   * **Hợp Tác với Các Chuyên Gia:** Làm việc với các nhà nghiên cứu và chuyên gia trong các lĩnh vực liên quan để cải thiện độ chính xác và ứng dụng của các mô hình phân tích, từ đó tạo ra các giải pháp hiệu quả hơn.

Các cải tiến này sẽ giúp nâng cao hiệu quả của dự án, cung cấp thông tin chính xác và giá trị hơn, và đáp ứng tốt hơn nhu cầu của người dùng và tổ chức.

5. Team Member Review and Comment

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| **NAME** | **REVIEW and COMMENT** |
| **Tào Chí Vỹ** | Thiết lập và quản lý hệ thống dữ liệu, bao gồm cấu hình Apache NiFi và Apache Kafka để truyền dữ liệu thời gian thực. Đảm bảo trích xuất, làm sạch, chuyển đổi và nạp dữ liệu vào MongoDB một cách hiệu quả. Phát triển và duy trì các bộ xử lý NiFi, các chủ đề Kafka và các bộ sưu tập MongoDB. Thực hiện các kỹ thuật làm sạch và tiền xử lý dữ liệu. |
| **Đặng Hoàng Toàn** | Phát triển và đánh giá mô hình phân tích cảm xúc. Tập trung vào khám phá dữ liệu, trích xuất đặc trưng, huấn luyện mô hình và đánh giá hiệu suất. Thực hiện tokenization, loại bỏ từ, áp dụng TF-IDF, huấn luyện mô hình hồi quy logistic và đánh giá hiệu suất mô hình bằng các chỉ số thích hợp. |
| **Trần Hoàng Anh Tú** | Thiết kế và phát triển ứng dụng web để hiển thị kết quả phân tích cảm xúc. Sử dụng Plotly và Dash để tạo các biểu đồ và bảng tương tác. Xây dựng giao diện người dùng, tích hợp các biểu đồ, đảm bảo thiết kế thân thiện với người dùng và triển khai các phản hồi ứng dụng cho dữ liệu động. |
|  |  |
|  |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |