Phân tích nhật ký thời gian thực Tổng quan về dự án

Phân tích Nhật ký là gì?

Quá trình đánh giá, hiểu và hiểu các tài liệu do máy tính tạo ra được gọi là nhật ký được gọi là phân tích nhật ký. Một loạt các công nghệ có thể lập trình, bao gồm thiết bị mạng, hệ điều hành, ứng dụng, v.v., tạo ra nhật ký. Nhật ký là một tập hợp các thông báo theo thứ tự thời gian mô tả những gì
đang diễn ra trong một hệ thống. Các tệp nhật ký có thể được phát tới bộ thu thập nhật ký qua mạng đang hoạt động hoặc được lưu trong các tệp để phân tích sau. Bất chấp điều đó, phân tích nhật ký là kỹ thuật tinh tế để đánh giá và giải thích các thông báo này nhằm hiểu rõ hơn về chức năng cơ bản của bất kỳ hệ thống nào. Phân tích nhật ký máy chủ web có thể cung cấp thông tin chi tiết quan trọng về mọi thứ, từ bảo mật đến dịch vụ khách hàng đến SEO. Thông tin được thu thập trong nhật ký máy chủ web có thể giúp bạn:

- Nỗ lực khắc phục sự cố mạng
- Phát triển và đảm bảo chất lượng
- Xác định và hiểu các vấn đề bảo mật
- Dịch vụ khách hàng
- Duy trì việc tuân thủ các chính sách của cả chính phủ và doanh nghiệp

Định dạng logfile phổ biến như sau:

remotehost rfc931 authuser [ngày] byte trạng thái "yêu cầu"

Đường ống dữ liệu:

Nó đề cập đến một hệ thống để di chuyển dữ liệu từ hệ thống này sang hệ thống khác. Dữ liệu có thể được chuyển đổi hoặc không và có thể được xử lý theo thời gian thực (hoặc truyền trực tuyến) thay vì theo đợt. Ngay từ việc trích xuất hoặc thu thập dữ liệu bằng nhiều công cụ khác nhau, lưu trữ dữ liệu thô, làm sạch, xác thực dữ liệu, chuyển đổi dữ liệu sang định dạng phù hợp với truy vấn, trực quan hóa KPI bao gồm cả Phối hợp của quy trình trên là đường dẫn dữ liệu.

Chương trình nghị sự của dự án là gì?

Chương trình nghị sự của dự án liên quan đến phân tích nhật ký thời gian thực với ứng dụng web trực quan hóa. Trước tiên, chúng tôi khởi chạy một phiên bản EC2 trên AWS và cài đặt Docker trong đó với các công cụ như Apache Spark, Apache NiFi, Apache Kafka, Jupyter Lab, Plotly và Dash. Sau đó, chúng tôi thực hiện tiền xử lý trên dữ liệu mẫu, phân tích dữ liệu thành các cột riêng lẻ, làm sạch dữ liệu và định dạng dấu thời gian. Tiếp theo là Khai thác tập dữ liệu nhật ký truy cập của NASA sử dụng Apache NiFi và Apache Kafka, tiếp theo là Chuyển đổi và Tải bằng Cassandra và HDFS và cuối cùng là Trực quan hóa nó bằng Python Plotly và Dash với việc sử dụng gọi lại ứng dụng biểu đồ và bảng.

Cách sử dụng Bộ dữ liệu:

Ở đây chúng ta sẽ sử dụng dữ liệu nhật ký truy cập của NASA theo các cách sau:

- Trích xuất: Trong quá trình trích xuất, tập dữ liệu đã tải xuống từ Kaggle được nhập bằng bộ xử lý và kết nối NiFi. Dữ liệu được truyền trực tuyến từ tệp dữ liệu bằng NiFi, sau đó là tạo chủ đề và xuất bản nhật ký bằng Apache Kafka.

- Chuyển đổi và tải: Trong quá trình chuyển đổi và tải, chúng tôi đọc dữ liệu từ Apache Kafka dưới dạng truyền Dataframe theo cách tạo lược đồ với việc trích xuất và làm sạch dữ liệu nhật ký và tải lên Cassandra cho lớp Tốc độ và HDFS cho lớp Batch. Sau đó, dữ liệu được trực quan hóa bằng cách sử dụng Plotly trong Dash.

Bài học rút ra

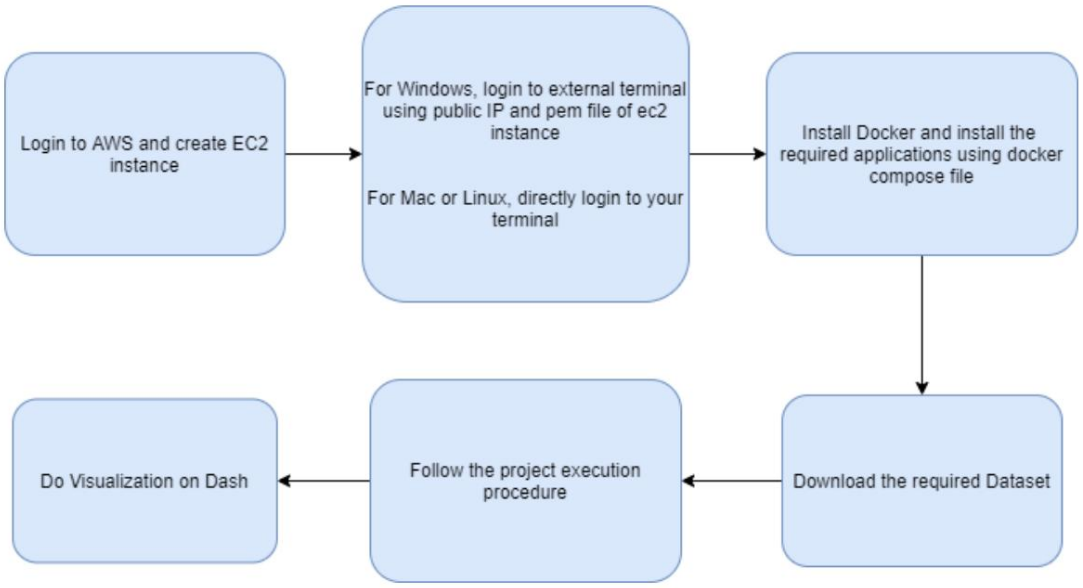
chính • Hiểu dự án và cách sử dụng Phiên bản AWS EC2 • Hiểu kiến thức cơ bản về Bộ chứa, phân tích nhật ký và ứng dụng của chúng • Trực quan hóa Kiến trúc hoàn chỉnh của hệ thống • Hiểu về Chuyển tiếp cổng • Giới thiệu về Docker

• Cách sử dụng docker-composer và khởi động tất cả các công cụ • Khám phá tập dữ liệu và định dạng nhật ký chung • Tìm hiểu Kiến trúc Lambda. • Cài đặt NiFi và sử dụng nó để nhập dữ liệu • Cài đặt Kafka và sử dụng nó để tạo chủ đề • Xuất bản nhật ký bằng NiFi • Tích hợp NiFi và Kafka • Cài đặt Spark và sử dụng nó để xử lý và làm sạch dữ liệu • Tích hợp Kafka và Spark • Đọc dữ liệu từ Kafka thông qua API truyền phát có cấu trúc Spark • Cài đặt và tạo không gian tên và bảng trong Cassandra • Tích hợp Spark và Cassandra • Tải dữ liệu liên tục trong Cassandra để có kết quả tổng hợp. • Tích hợp Cassandra và Plotly và Dash • Hiển thị luồng trực tiếp, kết quả Hàng giờ và Hàng ngày bằng Python Plotly và Dash

Phân tích dữ liệu:

- Từ trang web nhất định, dữ liệu được tải xuống có chứa dữ liệu nhật ký truy cập của NASA ở định dạng csv, chứa các thành phần khác nhau của nhật ký máy chủ web
- Bộ dữ liệu được xử lý, làm sạch và định dạng trường ngày giờ.
- Quá trình trích xuất được thực hiện bằng NiFi và Kafka, bằng cách truyền dữ liệu từ tệp nhật ký sử dụng NiFi và tạo chủ đề, xuất bản nhật ký bằng Kafka.
- Trong quá trình chuyển đổi và tải, lược đồ được xác định và dữ liệu được đọc từ Kafka dưới dạng Khung dữ liệu phát trực tuyến, lưu trữ trong Cassandra cho Đường dẫn nóng dưới Lớp tốc độ và trong Hadoop cho Đường dẫn lạnh dưới Lớp hàng loạt.
- Cuối cùng, dữ liệu được trực quan hóa bằng cách sử dụng các biểu đồ khác nhau theo cách Thời gian thực, Hàng giờ và Hàng ngày bằng cách sử dụng Plotly và Dash.

Quy trình làm việc của dự án:



Cấu trúc thư mục: