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

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 điệp theo thứ tự thời gian mô tả những gì
đang diễn ra trong một hệ thống. Tệp nhật ký có thể được phát tới trình thu thập nhật ký qua mạng đang hoạt động hoặc được lưu trong tệp để phân tích sau. Bất chấp điều đó, phân tích nhật ký là kỹ thuật tinh tế để đánh giá và diễn giải những 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 những hiểu biết quan trọng về mọi thứ, từ bảo mật, 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 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 phát 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 để truy vấn, trực quan hóa KPI bao gồm cả việc điều phối quy trình trên là đường dẫn dữ liệu.

Chương trình nghị sự của dự án là gì?

Chương trình làm việc của dự án liên quan đến phân tích nhật ký thời gian thực với ứng dụng web trực quan. Trước tiên, chúng tôi khởi chạy phiên bản EC2 trên AWS và cài đặt Docker trong đó bằng các công cụ như Apache Spark, Apache NiFi, Apache Kafka, Jupyter Lab, Plotly và Dash. Sau đó, chúng tôi thực hiện tiền xử lý 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 Trích xuất 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 Trực quan hóa nó bằng Python Plotly và Dash bằng cách sử dụng lệnh gọi lại ứng dụng bảng và biểu đồ.

Cách sử dụng Bộ dữ liệu:

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

- Trích xuất: Trong quá trình trích xuất, tập dữ liệu đã tải xuống từ Kaggle sẽ 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 phát Dataframe theo việc 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 hiển thị bằng Plotly trong Dash.

Những bài học chính •

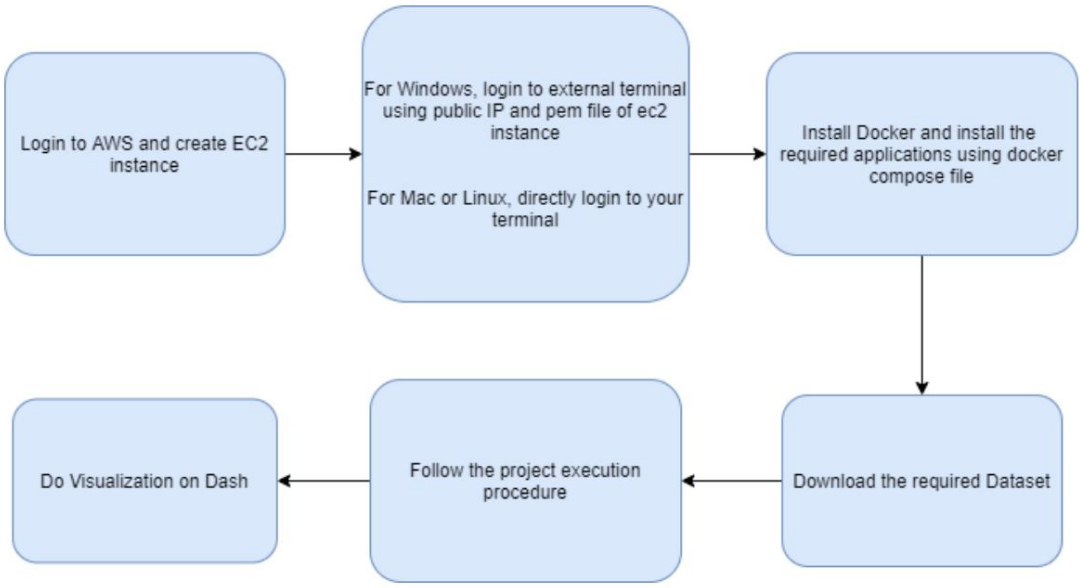
Hiểu dự án và cách sử dụng Phiên bản AWS EC2 • Tìm hiểu những kiến thức cơ bản về Bộ chứa, phân tích nhật ký và ứng dụng của chúng • Trực quan hóa Kiến trúc hoàn chỉnh của hệ thống • Tìm hiểu về Chuyển tiếp cổng • Giới thiệu về Docker

• Sử dụng docker-composer và khởi động tất cả các công cụ • Khám phá tập dữ liệu và định dạng nhật ký phổ biến • 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 phát trực tuyến 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 kết quả phát trực tiếp, Hàng giờ và Hàng ngày bằng Python Plotly và Dash

Phân tích dữ liệu:

- Từ trang web nhất định, dữ liệu được tải xuống 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
- Dataset được xử lý, làm sạch và định dạng trường datetime.
- 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 Dataframe phát trực tuyến, lưu trữ trong Cassandra cho Đường dẫn nóng trong Lớp tốc độ và trong Hadoop cho Đường dẫn lạnh trong Lớp hàng loạt.
- Cuối cùng, dữ liệu được hiển thị bằng cách sử dụng các biểu đồ khác nhau theo 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: