**-** **USE CASE -  
EXTRACT DATA FROM FORDGOBIKE, TRANSFORM AND STORE IT IN S3 AS CSV**

Thực hiện : Nguyễn Văn Tuấn - TuanNV90

Quản lý : Ứng Việt Anh - AnhUV

Chủ đề : Use Cases for Apache Nifi

**THUẬT NGỮ VÀ TỪ VIẾT TẮT**

S3 : Simple Storage Service

AWS : Amazon Web Services

**USE CASE 2:** **EXTRACT DATA FROM FORDGOBIKE, TRANSFORM AND** **STORE IT IN S3 AS CSV**

1. **Mô tả usecase**

* Trích xuất dữ liệu từ Ford GoBike (EXTRACT DATA FROM FORDGOBIKE):
* Mô tả: Trong bước này, hệ thống sẽ lấy dữ liệu từ API  
  <http://hamilton.socialbicycles.com/opendata/station_status.json> chứa thông tin về trạng thái của các trạm xe của Ford GoBike

A screenshot of a computer code

Description automatically generated

* Chuyển đổi dữ liệu và lưu trữ dữ liệu trên S3 dưới dạng .csv (TRANSFORM AND STORE IT IN S3 AS CSV):
* Mô tả: Sau khi các tệp đã được lấy từ thư mục tạm thời, bước tiếp theo là trích xuất thông tin về ID trạm và số lượng xe đạp có sẵn, sau đó chuyển đổi thông tin đó thành tệp CSV và thêm nó vào s3.

1. **Quy trình**

* Extract Data from Ford GoBike Api
* Do some processing (Apache Nifi): extract information about station IDs and the number of available bicycles, and then convert that information into a CSV file
* Put processed file into destination folder (S3)

1. **Chuẩn bị**

* Api: <http://hamilton.socialbicycles.com/opendata/station_status.json>
* Tài khoản Amazon Web Services (AWS)
* Apache Nifi đã dựng sẵn

1. **Thiết lập và cài đặt:**

Để Nifi có quyền đọc ghi trên S3, chúng ta cần phải thiết lập một cặp khoá truy cập (access key và secret access key). Để thiết lập cặp khoá truy cập, chúng ta cần tạo IAM user trong Amazon Web Services để lấy nó.

**Bước 1: Tạo IAM user**

Bước 1.1: Đăng nhập AWS

Bước 1.2: Truy cập vào trang quản lí các IAM user

* Lên thanh tìm kiếm của AWS console, gõ IAM và chọn “IAM”
* Chọn mục Users

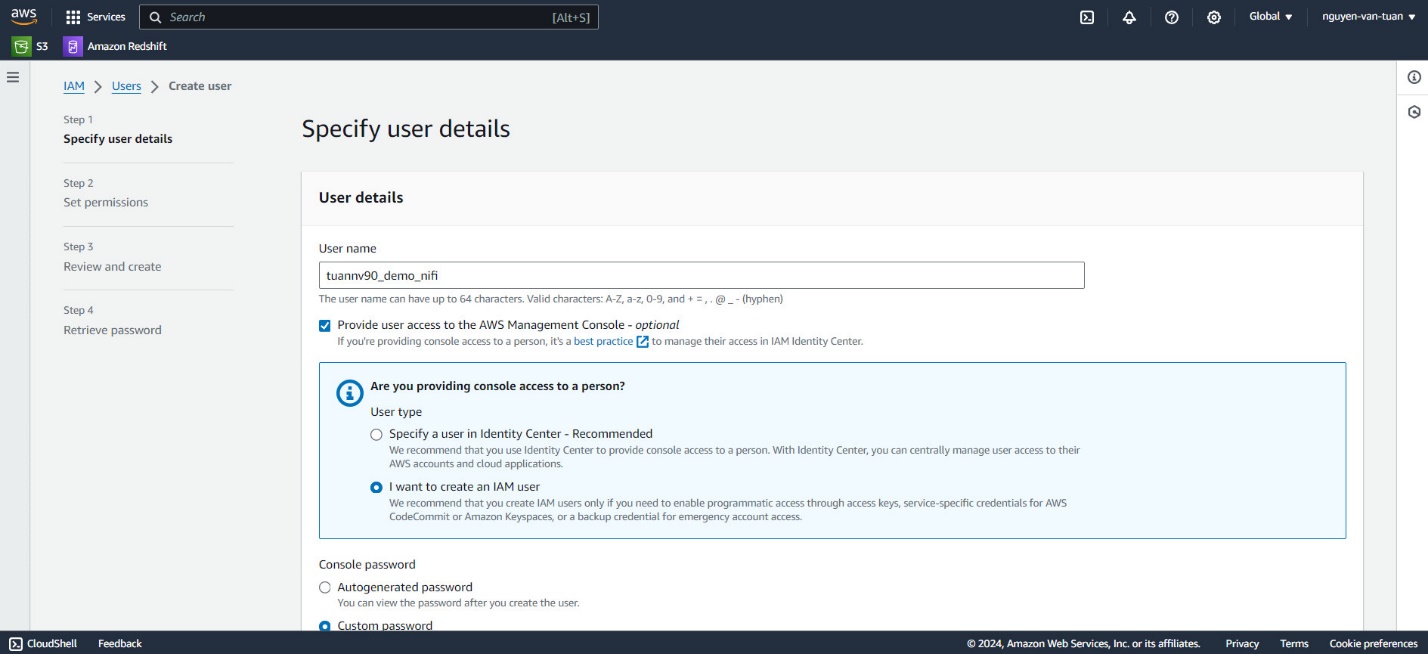
A screenshot of a computer

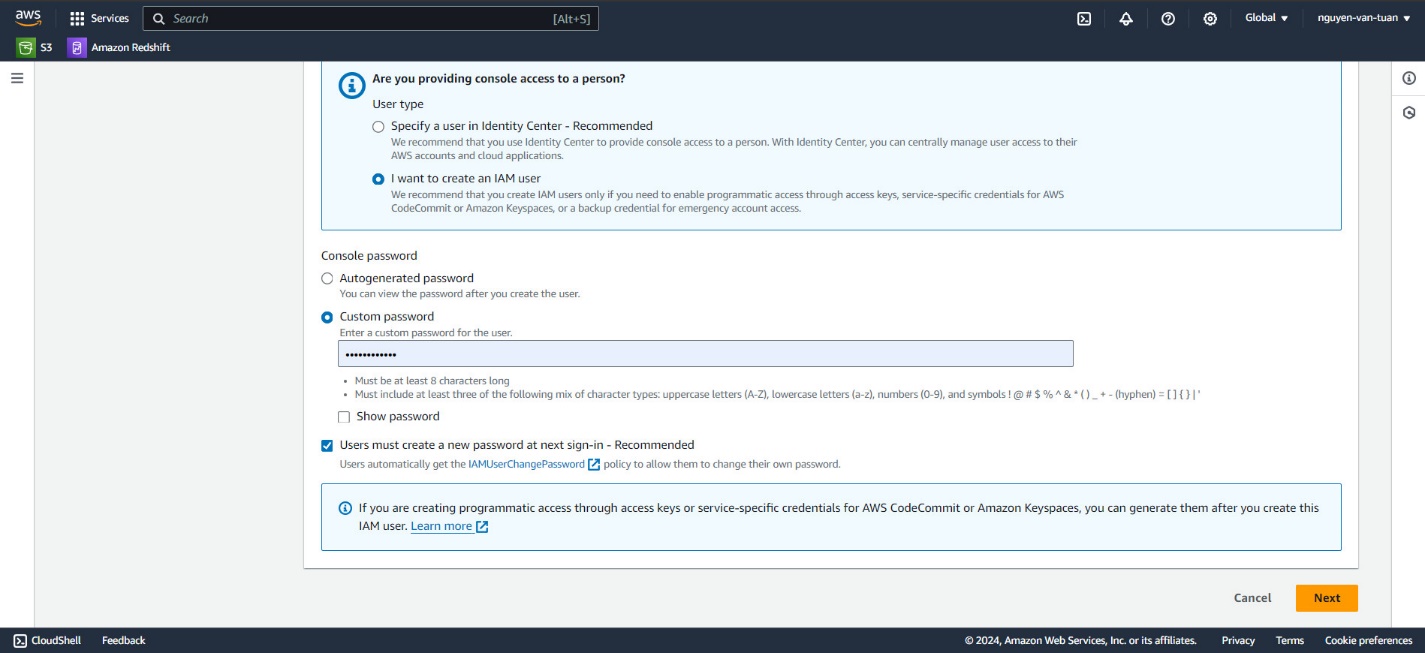
Description automatically generated

Bước 1.3: Chọn “Create user”

Bước 1.4: Thiết lập thông tin về IAM user

* User name: tên của User name
* Thiết lập thông tin như mẫu và chọn “Next”





Bước 1.5: Phân quyền cho IAM user

* Ấn Create group

A screenshot of a computer

Description automatically generated

* Đặt tên User group
* ở phần Permissions policies, tìm kiếm “s3” và chọn “AmazonS3FullAcess”
* Ấn Create user group

A screenshot of a computer

Description automatically generated

* Chọn User group vừa mới tạo và ấn “Next”

A screenshot of a computer

Description automatically generated

Bước 1.6: Chọn Create user

A screenshot of a computer

Description automatically generated

Bước 1.7: Chọn Download .csv file để tải file .csv chứa thông tin đăng nhập và chọn “Return to users list” để hoàn thành tạo IAM user.

A screenshot of a computer

Description automatically generated

**Bước 2: Tạo cặp khoá truy cập**

Bước 2.1: Chọn user vừa tạo trong danh sách các IAM user

A screenshot of a computer

Description automatically generated

Bước 2.2: Chọn Create access key

A screenshot of a computer

Description automatically generated

Bước 2.3: Chọn Third-party service và Next

A screenshot of a computer

Description automatically generated

Bước 2.4: Thêm Description tag và chọn Create access key

A screenshot of a computer

Description automatically generated

Bước 2.5: Chọn Download .csv file để lưu file csv chứa thông tin cặp khoá vì cặp khoá này chỉ có thể thấy được 1 lần duy nhất. Sau đó chọn Done để hoàn thành việc tạo cặp khoá truy cập

A screenshot of a computer

Description automatically generated

**Bước 3: Tạo nơi lưu trữ dữ liệu sau khi trích xuất trên S3**

Bước 3.1: Lên thanh tìm kiếm gõ S3 và chọn “S3”

Bước 3.2: Trong giao diện S3 chọn create bucket

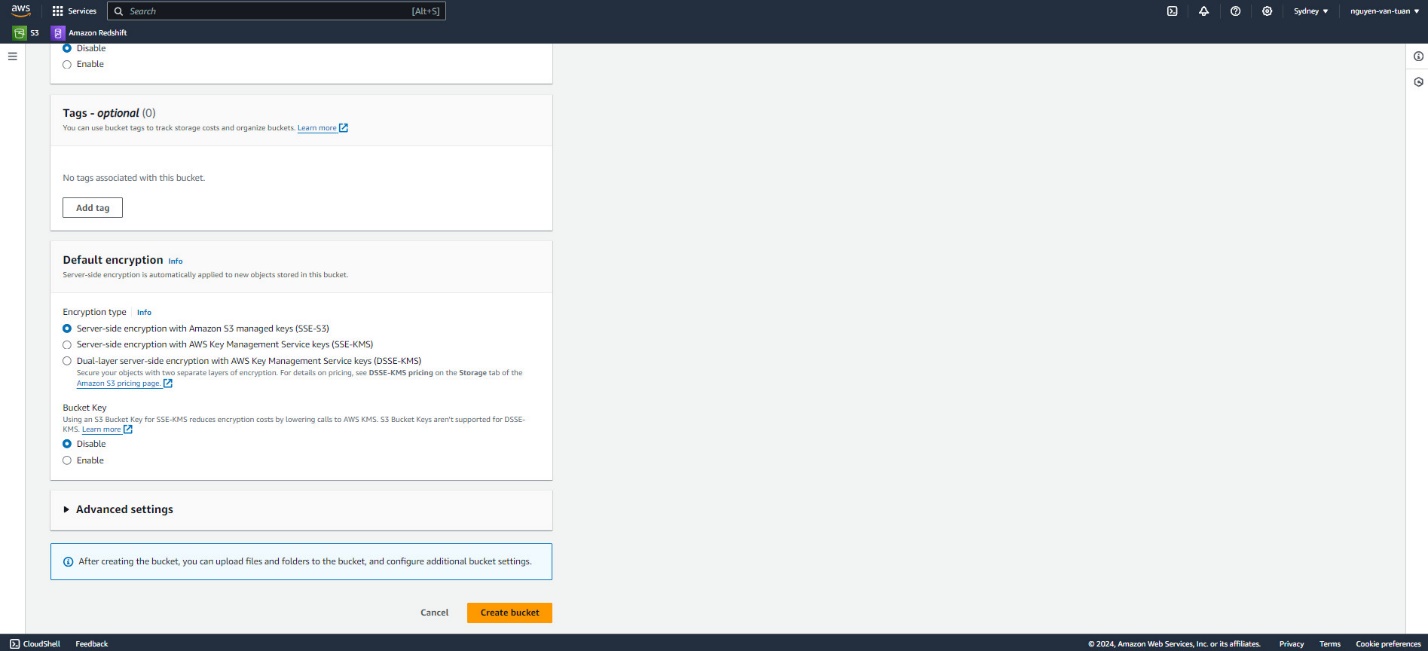
A screenshot of a computer

Description automatically generated

Bước 3.3: Điền “Bucket name” và thiết lập như mẫu, sau đó chọn create bucket để hoàn thành việc tạo bucket

A screenshot of a computer

Description automatically generated



**Bước 4: Triển khai luồng trên Apache Nifi**

Bước 4.1: Đầu tiên tạo một Process Group bằng việc kéo và thả vào giao diện, sau đó hãy đặt tên cho nó. Process Group này sẽ chứa các Processor và các kết nối giữa Processor của chúng ta, việc tạo Process Group giúp chúng ta quản lí dễ dàng hơn

A screenshot of a computer

Description automatically generated

Bước 4.2: Double-click vào Process Group vừa tạo, chúng ta sẽ kéo mục Processor để thêm các Processor và cấu hình nó như sau:

* InvokeHTTP
* Processor này dùng để đọc dữ liệu từ Ford GoBike Api và tạo ra flowfile trong Nifi

A screenshot of a computer

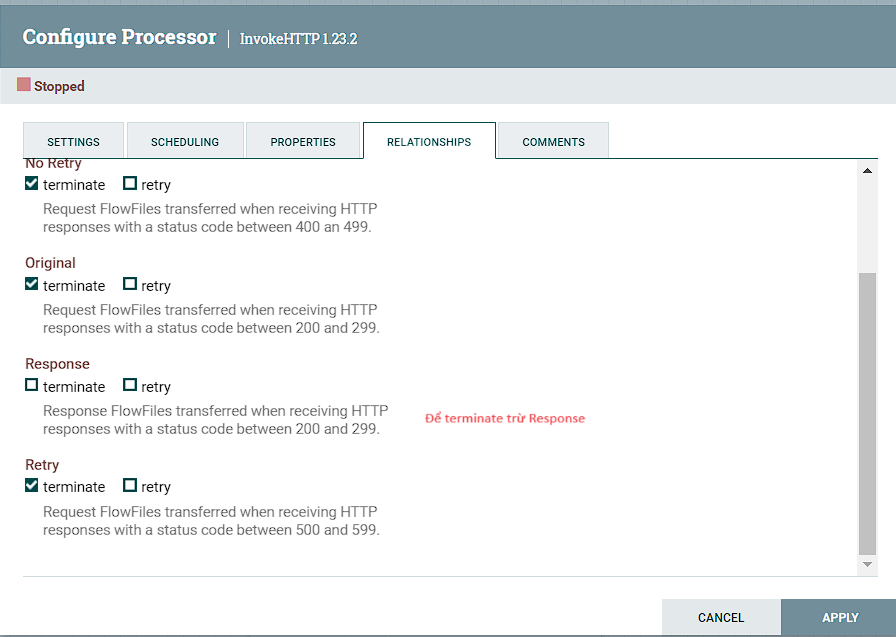
Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



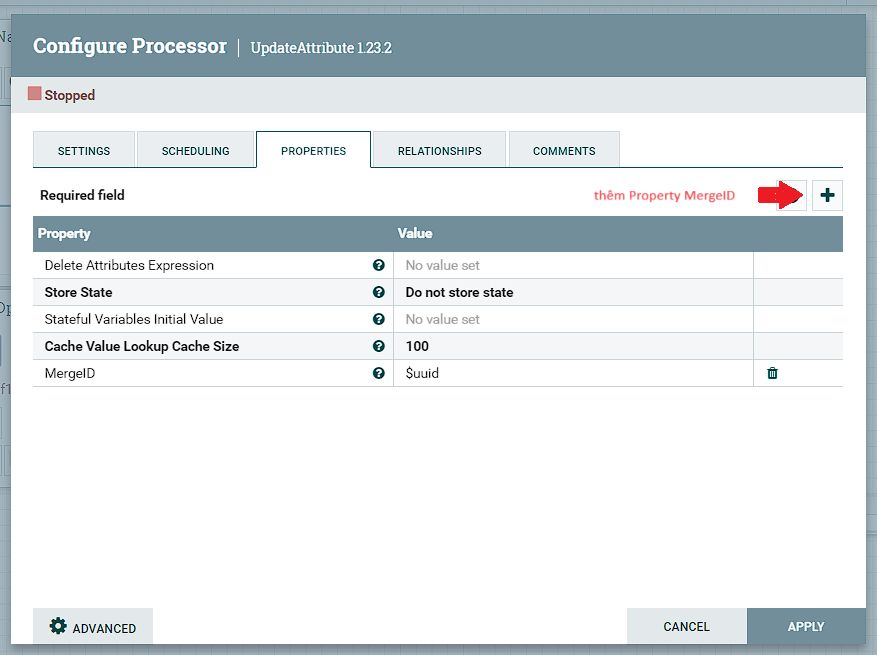
A screenshot of a computer program

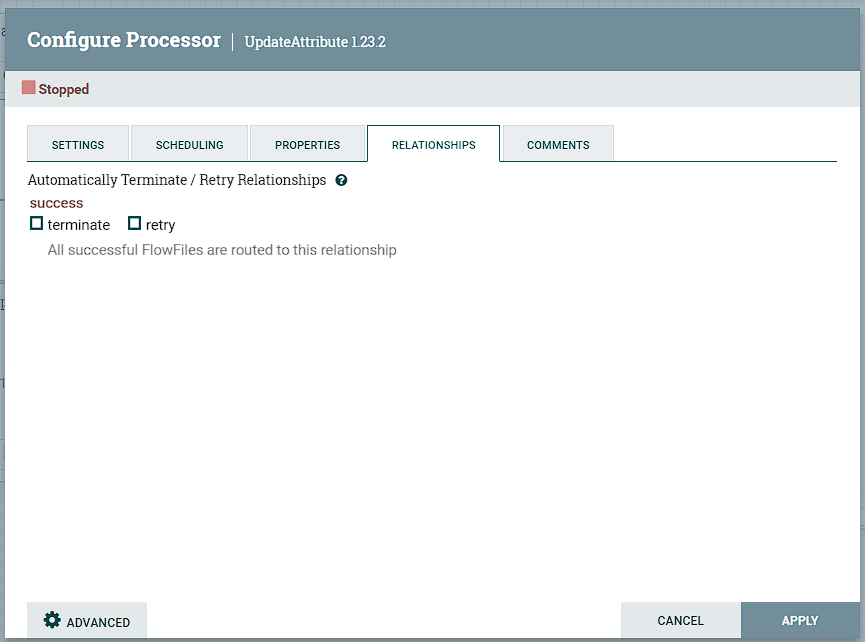
Description automatically generated

* UpdateAttribute
* sử dụng UpdateAttribute để thêm thuộc tính mergeID vào các flowfile, mục đích của mergeID sẽ là UUID duy nhất cho mỗi flowfile, sau khi tách ra để xử lí, sẽ dễ dàng hợp nhất lại các flowfile sau này.

A screenshot of a computer

Description automatically generated





* SplitJson
* Chúng ta tách các phần tử của stations thành các flowfile riêng để xử lí

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

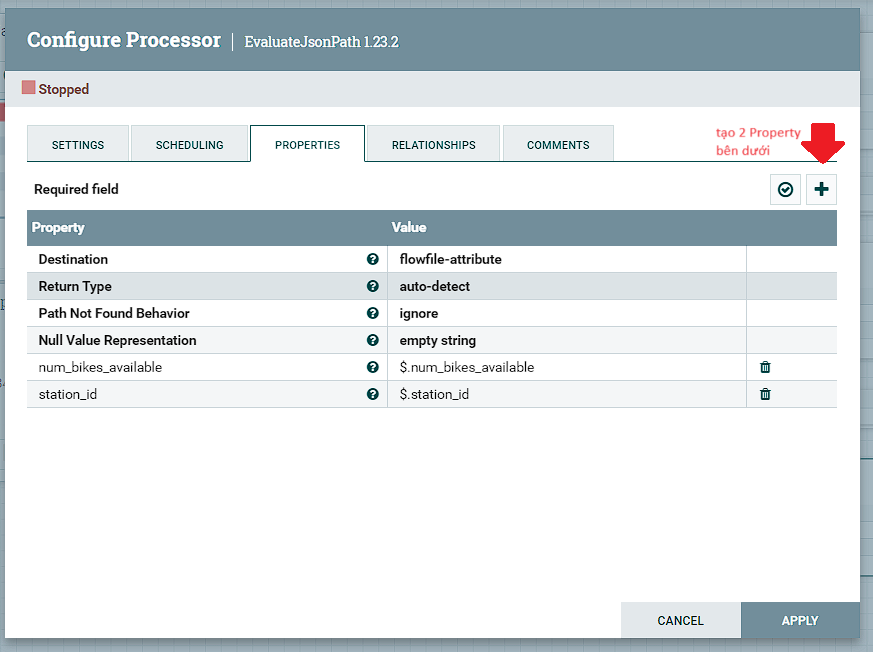
A screenshot of a computer

Description automatically generated

* EvaluateJsonPath
* Chúng ta sử dụng Processor này để trích xuất số lượng xe đạp có sẵn (num\_bikes\_available) và id của trạm (station\_id)

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

* ReplaceText
* thay thế tất cả nội dung của tệp luồng cụ thể đó theo định dạng ID trạm, số lượng xe đạp có sẵn.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* MergeContent
* Hợp nhất các flowfile theo “MergeId” sau khi đã xử lí và thêm Healder “stationid”, “verhicles available”

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* LogAttribute

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* UpdateAttribute
* Chuyển sang .csv

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* PutS3Project:
* Processor này dùng để tải lên (put) một FlowFile thành một đối tượng trong S3.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Chọn “Apply” để hoàn tất thiết lập

Bước 4.3: Kết nối các Processor để triển khai luồng

Chúng ta di chuột và nhấn kéo biểu tượng mũi tên từ processor này sang processor khác để tạo kết nối giữa chúng.

Đây là luồng chúng ta sẽ triển khai:

A diagram of a company

Description automatically generated with medium confidence

check dữ liệu trong bucket S3 đã tạo, chúng ta thấy dữ liệu đã được transform

=> hoàn thành dựng luồng

**Bước 5: Khởi chạy luồng**

Bước 5.1: Chuột phải vào khoảng trống trong giao diện và chọn “Start” để luồng bắt đầu hoạt động

A screenshot of a computer

Description automatically generated

Như vậy chúng đã triển khai thành công Usecase 2