Học phần Quản trị Dữ liệu lớn: Bài thực hành Apache Kafka (2)

Phạm Tiến Lâm, Đặng Văn Báu

Cài đặt thư viện:

1. Cài đặt **java** trên máy tính (điều kiện tiên quyết)
Link download: https://www.oracle.com/java/technologies/downloads

2. Cài đặt **Apache Kafka**
Link download: https://kafka.apache.org/downloads
Hướng dẫn cài đặt và sử dụng:
https://www.youtube.com/watch?v=BwYFuhVhshI&t=358s

3. Cài đặt **Jupyter notebook** hoặc **Anaconda**

4. Cài đặt **Kafka-Python**
pip install kafka-python
conda install -c conda-forge kafka-python

Chữa bài Lab6:

1. Khởi động máy chủ **Zookeeper**.

```
.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties
```

2. Khởi động máy chủ **Kafka**.

```
.\bin\windows\kafka-server-start.bat .\config\server.properties
```

3. Tạo **Topic** mới trong môi trường **Kafka**.

```
kafka-topics.bat --create --bootstrap-server localhost:9092 --replication-factor 1 --partition 1 --topic “topic_name”
```

4. Mở một '**producer console**' cho phép gửi các message tới **Topic** trong Kafka.

```
kafka-console-producer.bat --broker-list localhost:9092 --topic "topic_name"
```

5. Mở một "**consumer console**" và cho phép đọc tin nhắn từ **Topic**.

```
kafka-console-consumer.bat --topic "topic_name" --bootstrap-server localhost:9092 --from-beginning
```

6. Kiểm tra các **Topic** đã khởi tạo.

```
kafka-topics.bat --list --bootstrap-server localhost:9092
```

Activity 2: Bắt đầu với Python Apache Kafka

Tạo một jupyter notebook với tên "Producer"

- Khai báo thư viện sử dụng:

```
from kafka import KafkaProducer
from time import sleep
import json
```

- Sử dụng thư viện KafkaProducer để tạo một đối tượng producer, được kết nối đến một Kafka cluster thông qua các máy chủ

```
producer = KafkaProducer(bootstrap_servers= ['localhost:9092'], api_version = (0,10,1))
```

- Gửi 1 message tới topic "abcd" ("abcd" là Topic đã khởi tạo trước đó)

```
# topic name is abcd
producer.send('abcd', b'1')
```

- (*)Gửi một số nguyên lên Kafka với Topic 'abcd', sau khi chuyển đổi giá trị số này sang dạng bytes.

```
number_to_send = 20

# Chuyển đổi số nguyên thành bytes literal
number_bytes = str(number_to_send).encode('utf-8')

# Gửi thông điệp đến Kafka
producer.send('abcd', number_bytes)
```

- (**)Đọc một file dữ liệu "u.data" và gửi thông tin các dòng dữ liệu lên topic "abcd"

```
import csv

# Đọc file CSV và gửi từng dòng đến Kafka topic
with open('u.data', newline='') as csvfile:
    reader = csv.reader(csvfile)
    for row in reader:
        column_value = row[0]
        # Gửi từng cột của mỗi dòng đến Kafka topic
        for column_value in row:
            producer.send('abcd', value=column_value.encode('utf-8'))
            print(f"Sent value: {column_value}")

```

Tạo một jupyter notebook với tên "Consumer"

- Khai báo thư viện sử dụng

```
from kafka import KafkaConsumer
from time import sleep
import json
from json import loads
```

- Khởi tạo một đối tượng KafkaConsumer để kết nối đến một chủ đề Kafka ('abcd' trong trường hợp này) thông qua các máy chủ Kafka được chỉ định

```
consumer = KafkaConsumer('abcd',
                         bootstrap_servers= ['localhost:9092'],
                         api_version = (0,10))
```

- Đọc tất cả message nhận được từ Producer gửi tới Topic

```
for mess in consumer:
    print(mess.value)
```

- Từ phần (*) trong Activity 1: viết 1 đoạn mã lệnh trong Consumer tính bình phương số đó và in ra màn hình kết quả.

- Từ phần (**) trong Activity 1: viết 1 đoạn mã trong Consumer in ra dữ liệu nhận được từ Producer.

Bài tập:

Sử dụng thư viện **requests** và **BeautifulSoup** lấy thông tin từ trang web "https://vnexpress.net/apple-ra-mat-iphone-14-4508267.html"

Yêu cầu:

1. Khởi tạo 1 Producer: tìm ra nội dung thẻ <title> và truyền nội dung thẻ vào topic đã khởi tạo.
2. Khởi tạo một Consumer: in ra nội dung thẻ <title>
3. Khởi tạo 1 Producer: tìm ra nội dung tất cả các thẻ <a> và truyền nội dung thẻ vào topic đã khởi tạo.
4. Khởi tạo một Consumer: in ra nội dung thẻ <a>