1. Mục tiêu:

Mục tiêu của bài lab này là để làm quen với lập trình ứng dụng Spark, cụ thể hơn là sử dụng các hàm thao tác với RDD. RDD là một cấu trúc dữ liệu quan trọng của Spark.

2. Mô tả bài toán:

Tập dữ liệu được cung cấp là một log file của một hệ thống Content Delivery Network (CDN). Một hệ thống CDN đơn giản là một hệ thống caching dữ liệu nhằm giảm tải lưu lượng và chi phí truyền gói tin trong mạng. Từ đó, giảm độ trễ dịch vụ và tăng chất lượng trải nghiệm của khách hàng. Việc phân tích và hiểu rõ log file này giúp chúng ta hiểu rõ về giới hạn và chất lượng của hệ thống hiện tại, cũng như có thể đề xuất cho những hướng cải tiến khi cần mở rộng quy mô hệ thống.

Log file này chứa thông tin về lưu lượng, traffic, user request đi qua các server nằm của hệ thống trong 2 giờ. Log file dưới dạng thô, nên sẽ chứa các records sai định dạng hoặc thiếu thông tin. Trong bài lab này, nhiệm vụ của các bạn là vận dụng các hàm của Spark thao tác trên RDD để lọc, tiền xử lý, phân tích và tính toán một vài tham số thống kê trên log file này.

Đường dẫn trên server: *hdfs:///user/loclh/data/system_records.log*

Log Format: Mỗi dòng trong file tương ứng với một records. Mỗi record bao gồm 6 trường, gồm: độ trễ, địa chỉ IP của users, trạng thái cache, thời điểm hệ thống gửi trả response, tên nội dung của gói tin, và kích thước gói tin trả về.

Ví dụ, ta có một records sau:

1.001 58.187.29.147 HIT [02/Dec/2018:00:00:00 +0700] /live/prod_kplus_ns_hd/prod_kplus_ns_hd.isml/events(1541466558)/dash/prod_kplus_ns_hd-audio_vie=56000-49397873671168.dash 28401

1.001	Độ trễ	giây
58.187.29.147	Địa chỉ IP của users	
HIT	Trạng thái cache	Có 4 trạng thái: • HIT, HIT1: nội dung yêu cầu đã được cache bởi server. • MISS: nội dung yêu cầu chưa được cache • - : Trạng thái này xem như lỗi
[02/Dec/2018:00:00:00 +0700]	Thời điểm hệ thống gửi trả response	
/live/prod_kplus_ns_hd/prod_kplus_ns_hd.isml/events(1541466558)/dash/prod_kplus_ns_hd-audio_vie=56000-49397873671168.dash	Tên nội dung của gói tin	

28401	Kích thước gói tin trả về	Byte

3. Task: Lọc các records sai

Bước đầu tiên của quá trình phân tích dữ liệu luôn là bước lọc dữ liệu. Trong bài lab này, Một record được định nghĩa là sai nếu nó thỏa một trong các điều kiện sau:

- Records đó không chứa đúng 6 trường.
  (***Lưu ý:*** *các records đúng, khi sử dụng hàm split(" "), list trả về sẽ có* ***7*** *phần tử do có khoảng trắng " " trong trường "Thời điểm gửi yêu cầu")*
- Độ trễ phải là số dương (lớn hơn 0).
- Kích thước của gói tin phải là số dương
- Trạng thái cache không thể là dấu "-"

Yêu cầu:

- In ra tổng số records của cả log file và số records sai.
  (***Hint:*** Bạn có thể sử dụng một số hàm sau:
  - Spark: textFile(), map(), filter()
  - Python: split()

)
- Bây giờ sau khi đã lọc được danh sách các records đúng và các records sai. In ra top 10 records trễ nhất ở danh sách các records đúng. Các danh sách nên được sắp xếp theo thời gian response của server.
  (***Hint:*** Để có thể sắp xếp được thời gian, bạn nên chuyển đổi định dạng String của ngày tháng trong log thành định dạng timestamp (seconds hoặc milliseconds).
  - Spark: sortBy(), take()
  - Python: Để convert chuỗi ngày tháng thành định dạng timestamp, các bạn có thể tham khảo đoạn code sau

```
def convertStringToTimeStamp(inputString):
    from datetime import datetime
    timeWithoutZone = inputString.split(" ")[0] + "]"
    local_tz = pytz.timezone('Asia/Saigon')
    datetime_object = datetime.strptime(timeWithoutZone, '[%d/%b/%Y:%H:%M:%S]') #
    datetime_object = local_tz.localize(datetime_object).timestamp()

    return datetime_object
```