1. Mô tả luồng dữ liệu đi từ producer => kafka => consumer

* Message được producer gửi đến 1 broker với một topic cụ thể. Message được kafka xử lý và lưu vào các partition khác nhau.
* Một message thông thường sẽ sẽ có key để kafka điều hướng đúng chính xác partition để đảm bảo rằng message được đúng thứ tự.
* Consumer thực hiện subscribes một topic cụ thể để đọc message từ các topic

1. Dữ liệu lưu trữ trong kafka bị xoá đi khi nào?

* Dữ liệu lưu trữ trong kafka sẽ không bị xóa đi sau khi một message được xử lý thành công. Thông thường chúng ta sẽ config sau một khoảng thời gian nào đấy kafka sẽ xóa dữ liệu đi.

1. Khái niệm Consumer Group, tại một thời điểm có tối đa Consumer Group?

* Consumer group là một nhóm các consumer. Khi nhiều consumer subcribe tới một topic và đều thuộc về một Consumer Group này, mỗi consumer trong consumer group sẽ nhận về các message từ một tập các partition khác nhau trong topic. Consumer group read toàn bộ data của các partition và chia vào các consumer bên trong để xử lý. Mỗi consumer thuộc consumer group sẽ đọc toàn bộ data của một hoặc nhiều partition để đảm bảo message ordering. Không tồn tại việc nhiều consumer cùng đọc message từ một partition. Một consumer có thể nhận message từ nhiều partition. Nhưng một partition không thể gửi message cho nhiều consumer trong cùng consumer group.

1. Consumer trong kafka là pull (poll) dữ liệu từ các partition, khi 1 consumer khởi chạy, nó sẽ lấy dữ liệu như thế nào?

* Một consumer sau khi chạy nó sẽ connect đến một broker trong kafka cluster và request metadata (metadata chứa các thông tin về broker, topic, partition, offset của kafka cluster ) các thông tin này được zookeeper quản lý
* Sau khi có được thông tin của toàn bộ broker và metadata, consumer sẽ quyết định connect đến broker nào để lấy data
* Consumer sẽ liên tục gửi request đến kafka để lấy về các message mới.
* Khi consumer xử lý xong message, chúng ta commit giá trị offset lên kafka, các giá trị này sẽ lưu tại Kafka topic có tên là \_\_consumer\_offsets.

1. So sánh kafka & rabbitmq dựa trên kiến thức của bản thân

* Kafka lưu trữ dữ liệu trên ổ cứng còn rabbit lưu trữ trên ram
* Với kafka consumer phải chủ động pull message về còn rabbit thì sẽ push message cho consumer
* Kafka quản lý message theo thứ tự còn rabbit thì không
* Kafka có giới hạn consumer còn rabbit thì không
* Kafka message sẽ được xóa đi trong thời gian nhất định còn rabbit message sẽ bị xóa sau khi tất cả consumer nhận được message từ topic đã subscriber

1. Điều gì xảy ra khi số lượng consumer của một consumer group lớn hơn số partition?

* Nếu số lượng consumer trong consumer group lớn hơn số lượng partition thì những consumer dư thừa có trạng thái inactive - không nhận bất kì message nào từ topic.
* Trong trường hợp một active consumer gặp vấn đề và không thể tiếp tục hoạt động, một trong những inactive consumer còn lại được đẩy lên thay thế và tiếp tục công việc ngay lập tức.
* Nếu không có inactive consumer nào, message sẽ được route tới một active consumer bất kì khác.

1. Nói kafka xử lý được các bài toán liên quan đến thứ tự, vì sao lại thế?

* Vì việc đọc các message trong partition được diễn ra tuần tự có nghĩa là consumer không thể đọc message offset=3 khi chưa đọc message offset=2.