**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**——ﻩ¤🙞🙜¤ﻩ——**

A blue and white logo

Description automatically generated

**BÁO CÁO BÀI TẬP 4  
Spark Streaming**

**Bộ môn: NHẬP MÔN DỮ LIỆU LỚN**

**Giảng viên LT: Nguyễn Ngọc Thảo**

***TP.Hồ Chí Minh, 03/2025***

**MỤC LỤC**

[PHẦN 1: GIỚI THIỆU SINH VIÊN 2](#_Toc185529643)

[PHẦN 2: NỘI DUNG. 2](#_Toc185529644)

[Bài 2.1. 2](#_Toc185529645)

[Bài 2.2. 5](#_Toc185529646)

# 

PHẦN I. GIỚI THIỆU NHÓM

|  |  |  |
| --- | --- | --- |
| MSSV | Họ và tên | Công việc |
|  |  |  |
|  |  |  |
|  |  |  |
| 22120335 | Trương Tâm Thành | Bài 1: Extract |

PHẦN II. NỘI DUNG

1/ Extract

A screenshot of a computer

AI-generated content may be incorrect.

A black background with blue text

AI-generated content may be incorrect.

* Dùng docker-compose để tạo Kafka.

A screen shot of a computer

AI-generated content may be incorrect.

* Dùng cơ chế Aligned Scheduling Algorithm, loại bỏ hiện tượng lệch thời gian tích lũy => Message được gửi chính xác tại các mốc 0ms, 100ms, 200ms,...



* Làm tròn timestamp về các bucket 100ms để xử lý late data

A screen shot of a computer code

AI-generated content may be incorrect.

A computer code on a black background

AI-generated content may be incorrect.

* Ưu tiên dùng giá mới nhất. Nếu không có giá mới, sử dụng giá cuối cùng hợp lệ. Đảm bảo luôn có dữ liệu để gửi.



A screen shot of a computer

AI-generated content may be incorrect.

* Dùng đa luồng fetch dữ liệu để giảm ảnh hưởng độ trễ mạng.

A screen shot of a computer

AI-generated content may be incorrect.

* Điều kiện cập nhật: giá thay đổi hoặc đã gửi 5 message liên tiếp cùng giá trị hoặc dữ liệu cũ hơn 500ms

A computer screen with white text

AI-generated content may be incorrect.

* Ngoài event\_time thì thêm data\_age\_ms để đánh giá độ trễ

A black screen with white text

AI-generated content may be incorrect.

* Check symbol và price có trong data hay không

A black background with white text

AI-generated content may be incorrect.

* Bản ghi sau khi chuẩn hóa được gửi tới Kafka topic có tên chính xác là btc-price.

A screenshot of a computer screen

AI-generated content may be incorrect.

* Đã gửi đúng định dạng.

A screen shot of a computer

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

* Kiểm tra trong Kafka bằng cách shell vô container Kafka ở Docker, Kafka đã nhận đúng.

2/ Transform

3/ Load

4/ Bonus