



HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Xây dựng hệ thống hồ dữ liệu phân tích dữ liệu chuyển bay sử dụng các thành phần của hệ sinh thái Hadoop

Sinh viên thực hiện: Trần Ngọc Bảo
Mã số sinh viên: 20215529
Giảng viên hướng dẫn: TS. Trần Việt Trung

ONE LOVE. ONE FUTURE.



Nội dung

1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
3. Thiết kế hệ thống
4. Triển khai hệ thống
5. Kết quả thực nghiệm
6. Kết luận





Nội dung

1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
3. Thiết kế hệ thống
4. Triển khai hệ thống
5. Kết quả thực nghiệm
6. Kết luận

1. Giới thiệu đề tài

- Dữ liệu chuyến bay ngày càng đa dạng và khổng lồ, bao gồm lịch trình, tình trạng bay, thời gian trễ, thời tiết và nhiều yếu tố khác.
- Đòi hỏi doanh nghiệp xây dựng hệ thống phân tích hiệu quả dựa trên mô hình triển khai và hệ sinh thái công nghệ phù hợp để tối ưu hóa quản lý, hỗ trợ quyết định và nâng cao cạnh tranh.



Nội dung

1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
3. Thiết kế hệ thống
4. Triển khai hệ thống
5. Kết quả thực nghiệm
6. Kết luận

2. Mục tiêu và giải pháp

- Đề án tập trung xây dựng hệ thống phân tích dữ liệu chuyển bay dựa trên mô hình hồ dữ liệu kết hợp hệ sinh thái Hadoop, tận dụng khả năng lưu trữ linh hoạt và xử lý dữ liệu lớn trên hệ thống phân tán với chi phí thấp.
- Hệ thống hỗ trợ thu thập, xử lý, phân tích và trực quan hóa dữ liệu nhằm tối ưu hiệu suất, giảm chi phí vận hành và hỗ trợ doanh nghiệp hàng không ra quyết định hiệu quả.

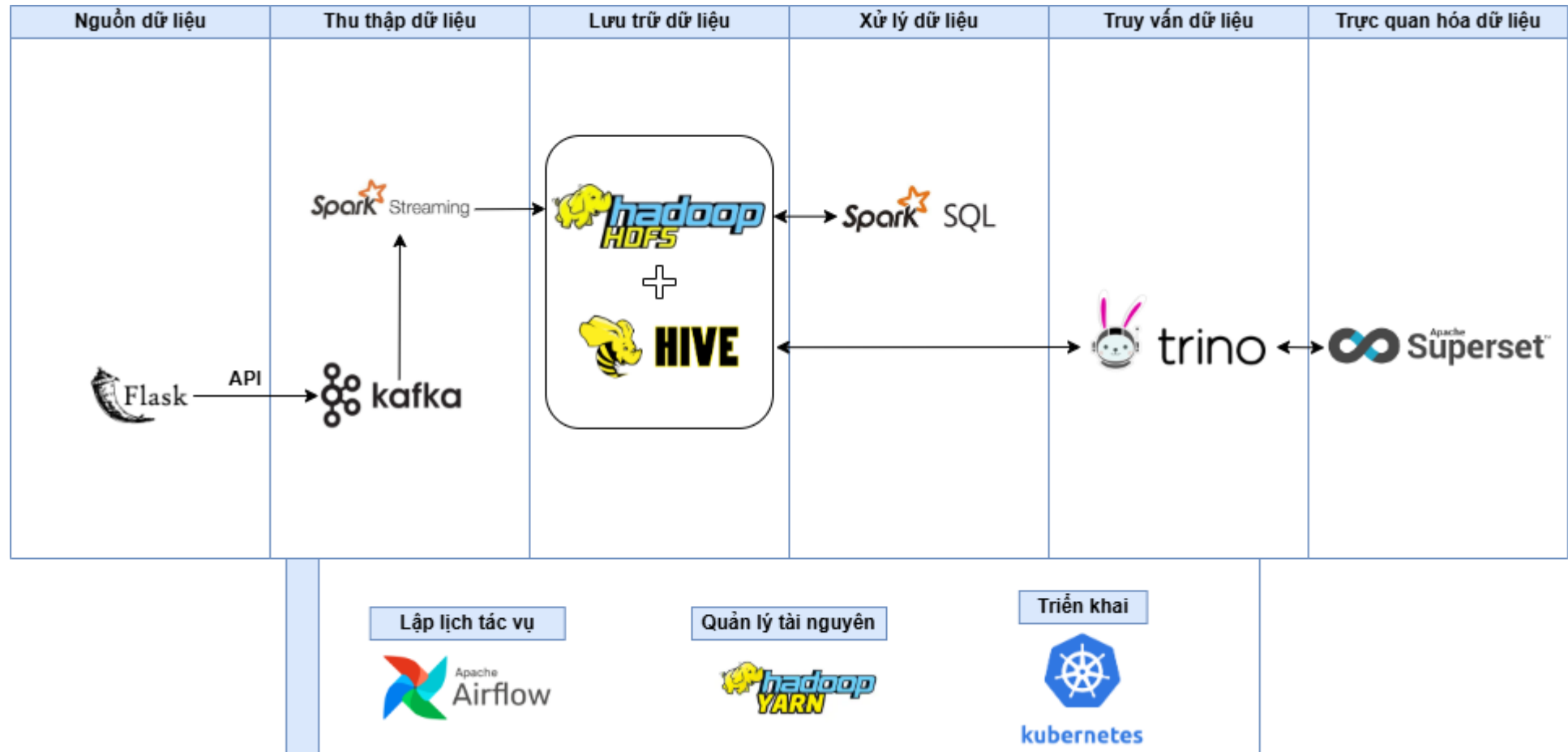


Nội dung

1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
- 3. Thiết kế hệ thống**
4. Triển khai hệ thống
5. Kết quả thực nghiệm
6. Kết luận

3. Thiết kế hệ thống

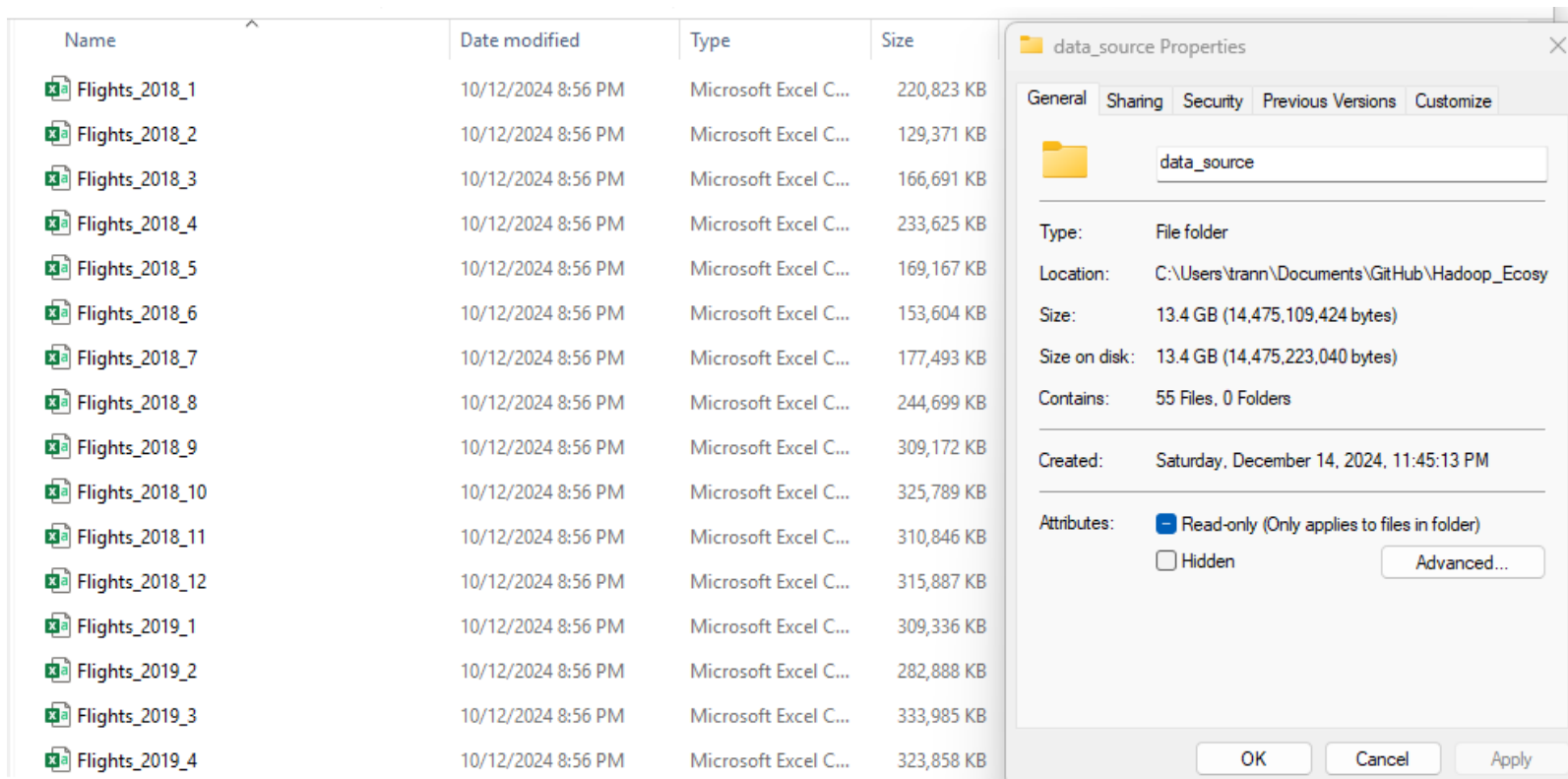
Tổng quan hệ thống



3. Thiết kế hệ thống (Mô-đun lưu trữ dữ liệu)

Trong hệ thống có 3 nơi lưu trữ dữ liệu đó là:

(i) Dữ liệu nguồn tại máy chủ Flask cung cấp API: dữ liệu được lưu trữ dưới dạng các file CSV có tên dưới dạng: "Flights_{year}_{month}", trong đó "year" là năm và "month" là tháng của file dữ liệu.



Name	Date modified	Type	Size
Flights_2018_1	10/12/2024 8:56 PM	Microsoft Excel C...	220,823 KB
Flights_2018_2	10/12/2024 8:56 PM	Microsoft Excel C...	129,371 KB
Flights_2018_3	10/12/2024 8:56 PM	Microsoft Excel C...	166,691 KB
Flights_2018_4	10/12/2024 8:56 PM	Microsoft Excel C...	233,625 KB
Flights_2018_5	10/12/2024 8:56 PM	Microsoft Excel C...	169,167 KB
Flights_2018_6	10/12/2024 8:56 PM	Microsoft Excel C...	153,604 KB
Flights_2018_7	10/12/2024 8:56 PM	Microsoft Excel C...	177,493 KB
Flights_2018_8	10/12/2024 8:56 PM	Microsoft Excel C...	244,699 KB
Flights_2018_9	10/12/2024 8:56 PM	Microsoft Excel C...	309,172 KB
Flights_2018_10	10/12/2024 8:56 PM	Microsoft Excel C...	325,789 KB
Flights_2018_11	10/12/2024 8:56 PM	Microsoft Excel C...	310,846 KB
Flights_2018_12	10/12/2024 8:56 PM	Microsoft Excel C...	315,887 KB
Flights_2019_1	10/12/2024 8:56 PM	Microsoft Excel C...	309,336 KB
Flights_2019_2	10/12/2024 8:56 PM	Microsoft Excel C...	282,888 KB
Flights_2019_3	10/12/2024 8:56 PM	Microsoft Excel C...	333,985 KB
Flights_2019_4	10/12/2024 8:56 PM	Microsoft Excel C...	323,858 KB

data_source Properties

General Sharing Security Previous Versions Customize

data_source

Type: File folder

Location: C:\Users\trann\Documents\GitHub\Hadoop_Ecosy

Size: 13.4 GB (14,475,109,424 bytes)

Size on disk: 13.4 GB (14,475,223,040 bytes)

Contains: 55 Files, 0 Folders

Created: Saturday, December 14, 2024, 11:45:13 PM

Attributes: ☒ Read-only (Only applies to files in folder)
☐ Hidden

Advanced...

OK Cancel Apply

3. Thiết kế hệ thống (Mô-đun lưu trữ dữ liệu)

(ii) Hàng đợi thông điệp tại cụm Kafka: được chia theo các topic, partition, offset. Trong đó topic được chia theo các năm, cụm Kafka sẽ có 3 partition và được nhân bản 3 lần để đảm bảo dữ liệu được thu thập và lưu trữ an toàn.

(iii) Dữ liệu lưu trên HDFS: là nơi lưu trữ dữ liệu chính của hệ thống. HDFS lưu trữ dữ liệu sau quá trình tiền xử lý Spark Streaming, dữ liệu sau quá trình xử lý theo lô của Spark SQL và các bảng dữ liệu được tạo ra sau quá trình truy vấn dữ liệu bằng Trino.

3. Thiết kế hệ thống (Mô-đun xử lý dữ liệu)

Xử lý dữ liệu thời gian thực:

Sử dụng Spark Streaming, dữ liệu thời gian thực từ Kafka được xử lý và lưu trữ vào HDFS.

```
df = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "kafka:9092") \
    .option("subscribe", f"flight_data_{year}") \
    .load()

json_df = df.selectExpr("CAST(key AS STRING) as msg_key", "CAST(value AS STRING) as msg_value")

json_expanded_df = json_df.withColumn("msg_value", from_json(json_df["msg_value"], json_schema)).select("msg_value.*")

writing_df = json_expanded_df.writeStream \
    .format("parquet") \
    .option("format", "append") \
    .option("path", "hdfs://hadoop-hadoop-hdfs-nn:9000/staging/" + str(year) + "/" + str(month)) \
    .option("checkpointLocation", "hdfs://hadoop-hadoop-hdfs-nn:9000/tmp/" + str(year) + "/" + str(month)) \
    .outputMode("append") \
    .start()

def stop_query():
    writing_df.stop()

timer = threading.Timer(24 * 60 * 60, stop_query)
timer.start()

writing_df.awaitTermination()
```

3. Thiết kế hệ thống (Mô-đun xử lý dữ liệu)

Xử lý dữ liệu theo lô:

Sử dụng Spark SQL thực hiện các công việc xử lý dữ liệu sau:

- (i) Loại bỏ các bản ghi trùng lặp
- (ii) Xử lý các cột dữ liệu liên quan đến thời gian
- (iii) Chuyển đổi kiểu dữ liệu
- (iv) Kết hợp với Hive tạo bảng lưu trữ trên HDFS




Nội dung

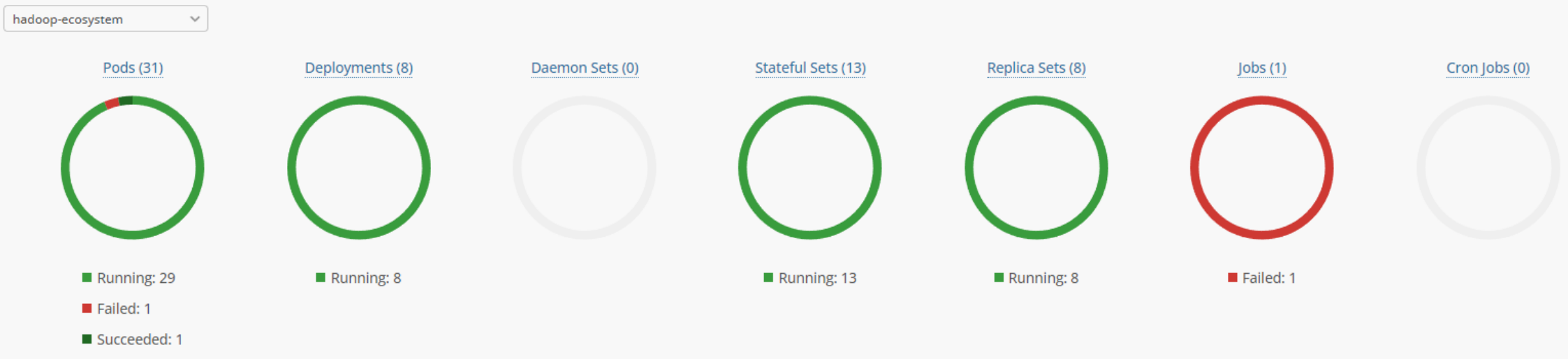
1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
3. Thiết kế hệ thống
- 4. Triển khai hệ thống**
5. Kết quả thực nghiệm
6. Kết luận

4. Triển khai hệ thống

Cụm Kubernetes của hệ thống

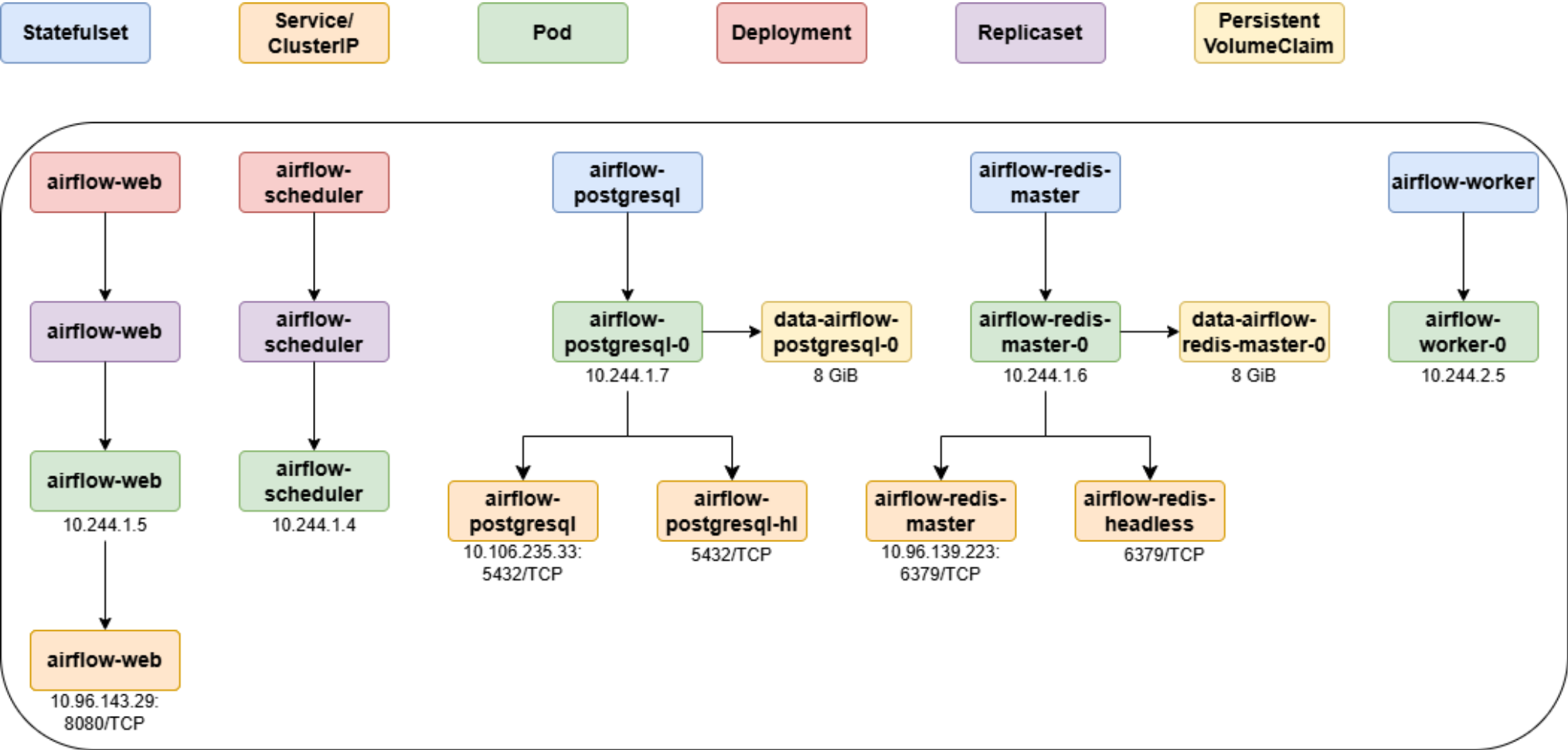
<input type="checkbox"/> Name	 CPU	Memory	Disk	Taint	Roles	Version	Age	Conditions
<input type="checkbox"/> hadoop-ecosystem	<div><div></div></div>	<div><div></div></div>	<div><div></div></div>	0	control-plane	v1.31.0	8h	Ready
<input type="checkbox"/> hadoop-ecosystem-m02	<div><div></div></div>	<div><div></div></div>	<div><div></div></div>	0	worker	v1.31.0	8h	Ready
<input type="checkbox"/> hadoop-ecosystem-m03	<div><div></div></div>	<div><div></div></div>	<div><div></div></div>	0	worker	v1.31.0	8h	Ready

Tổng quan trạng thái của hệ thống



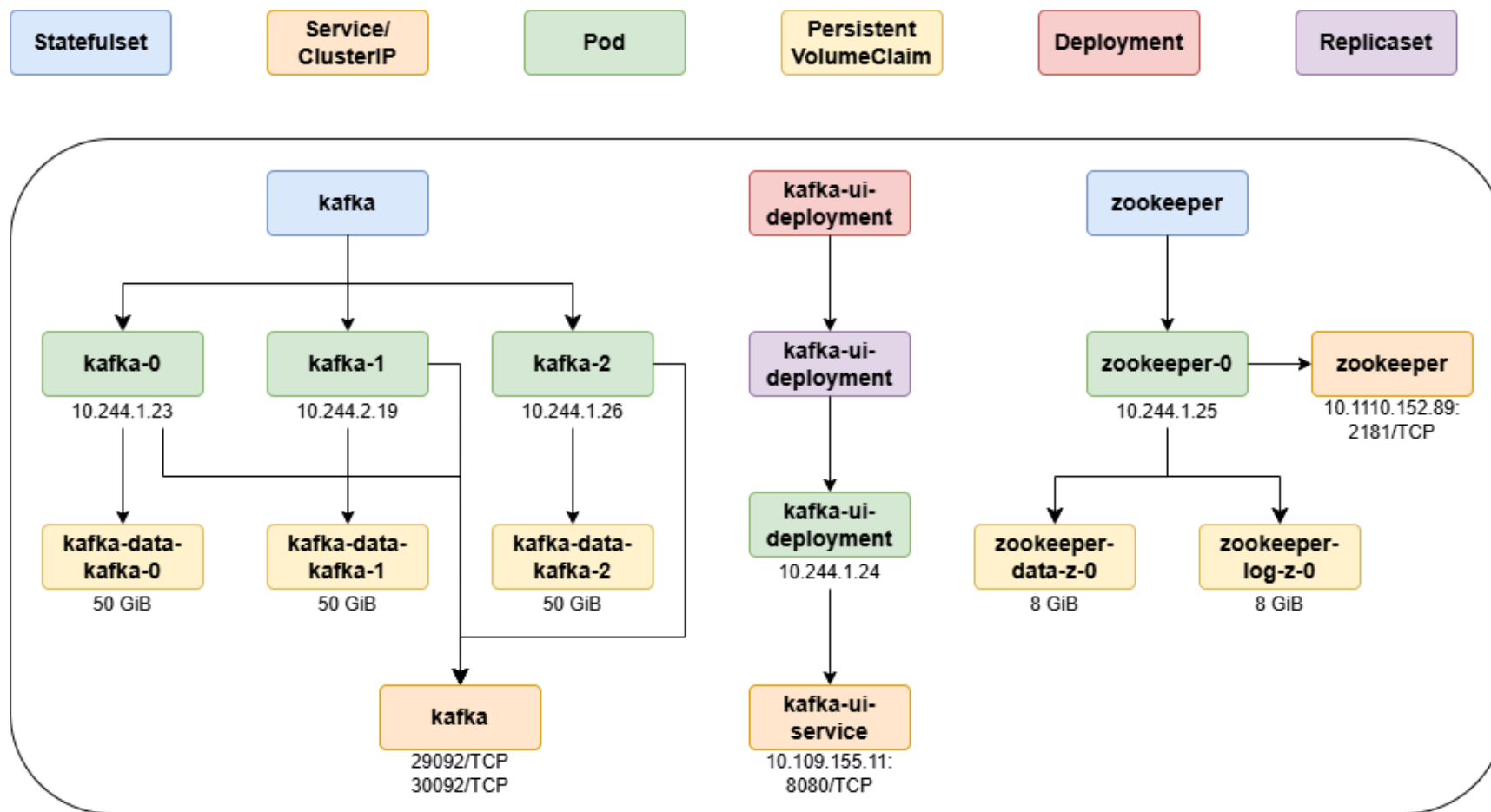
4. Triển khai hệ thống

Triển khai cụm Airflow



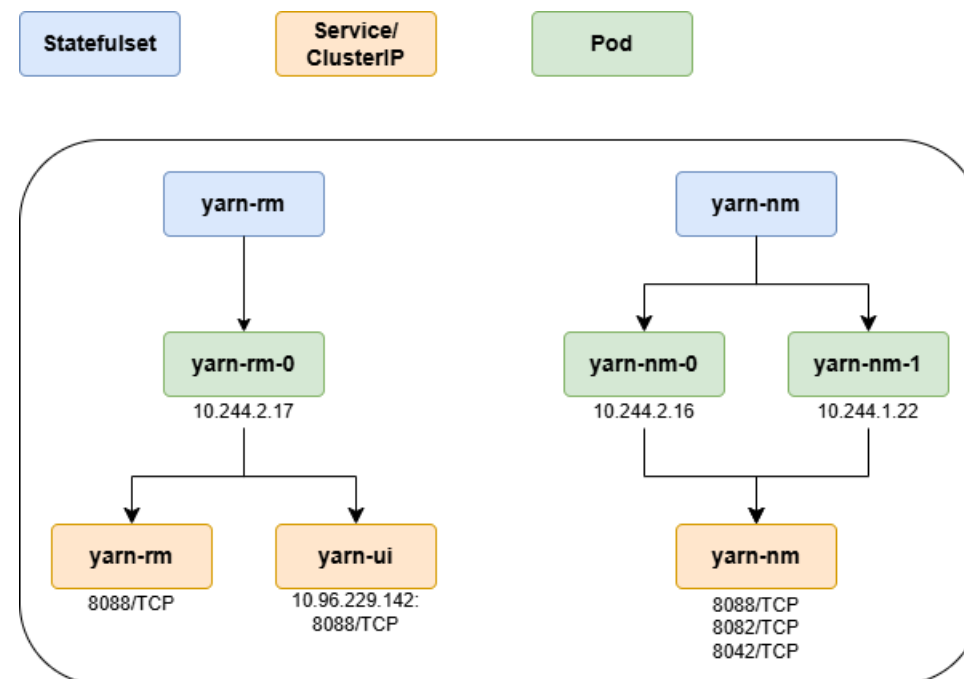
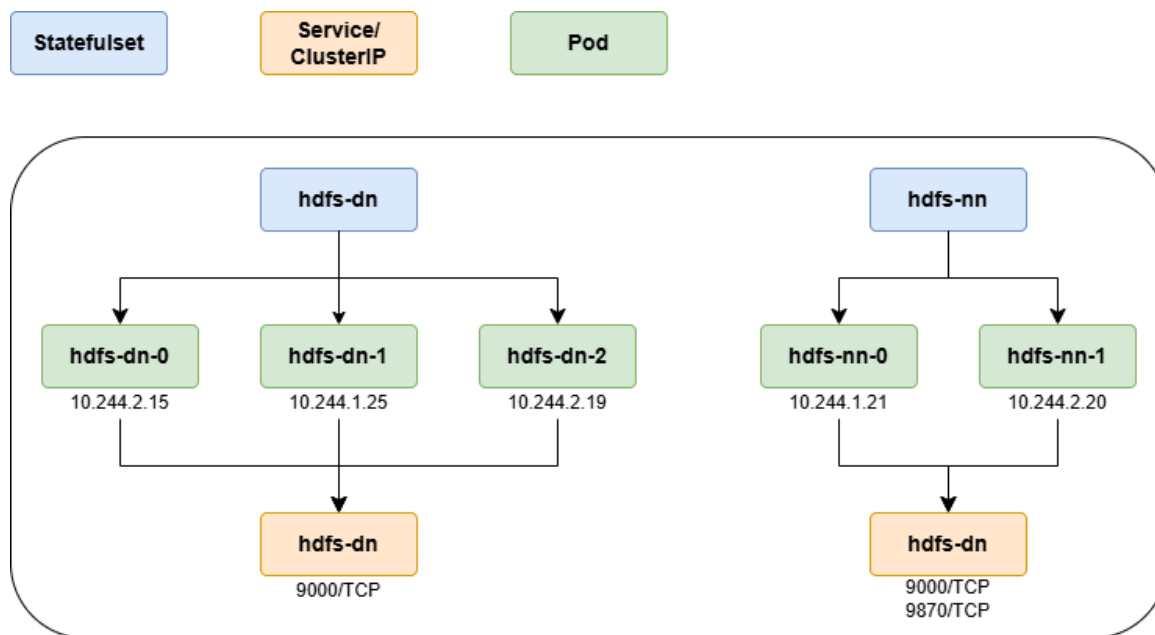
4. Triển khai hệ thống

Triển khai cụm Kafka



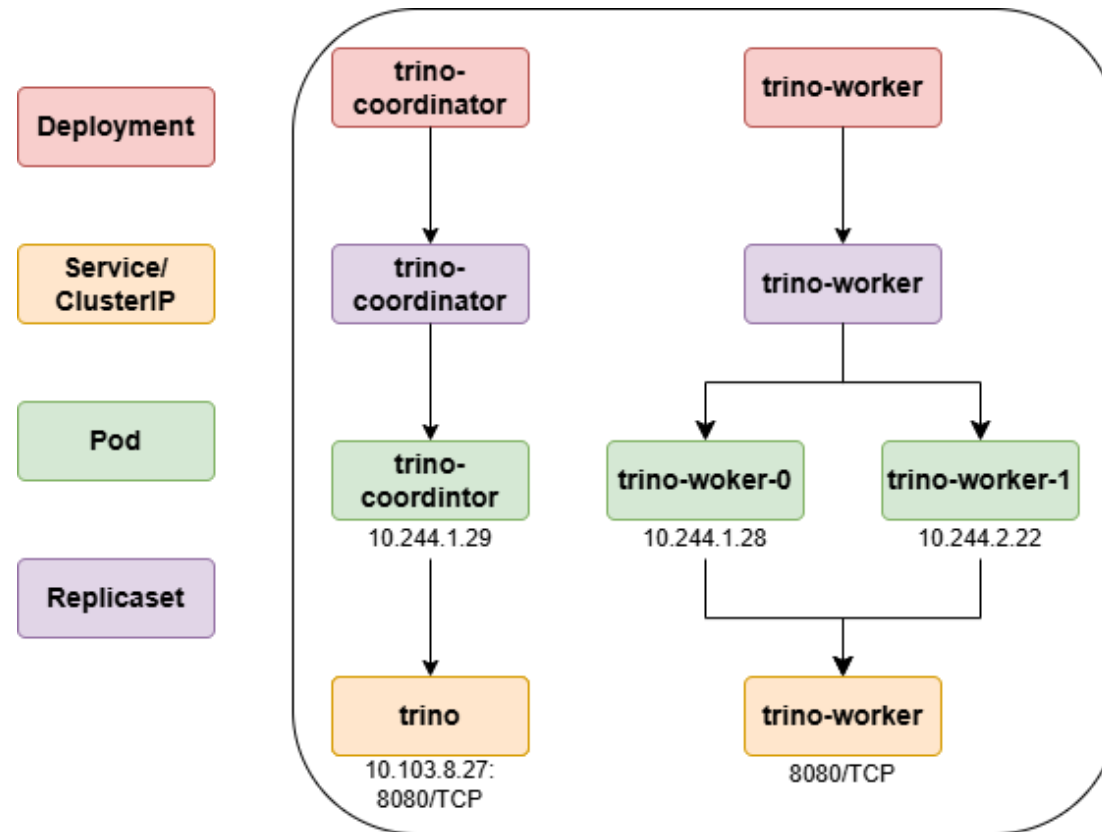
4. Triển khai hệ thống

Triển khai cụm HDFS và cụm YARN



4. Triển khai hệ thống

Triển khai cụm Trino



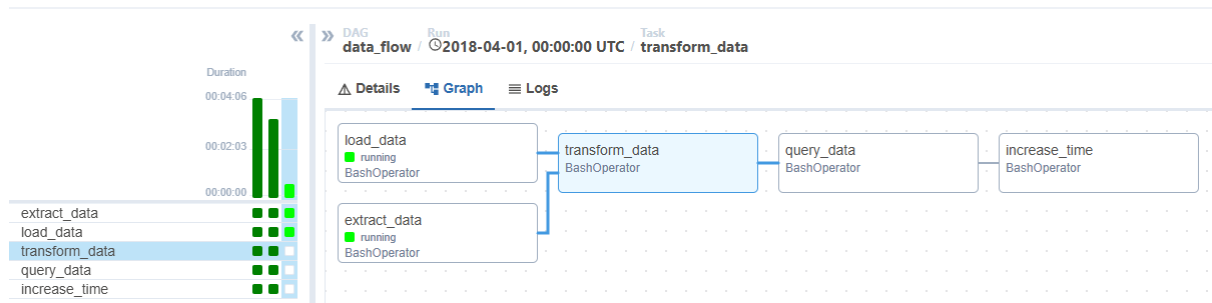


Nội dung

1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
3. Thiết kế hệ thống
4. Triển khai hệ thống
- 5. Kết quả thực nghiệm**
6. Kết luận

5. Kết quả thực nghiệm

Tính tự động hóa



5. Kết quả thực nghiệm

Tính khả mở

```
C:\Users\trann>kubect1 scale --replicas=2 statefulset/zookeeper
statefulset.apps/zookeeper scaled

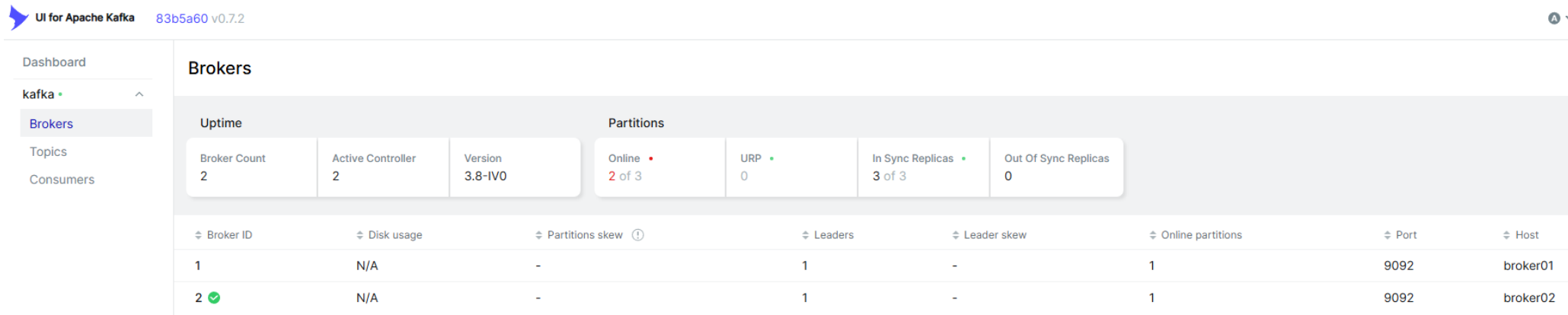
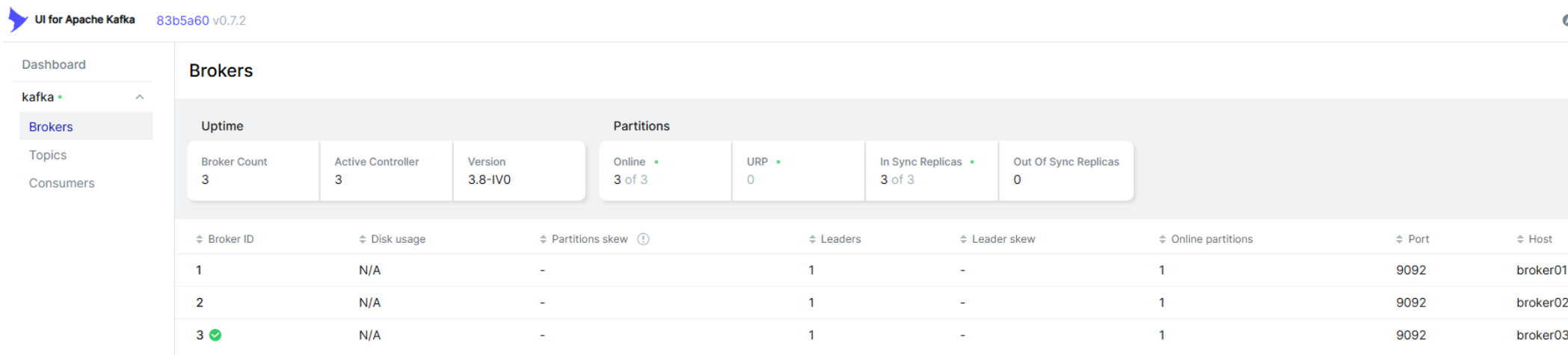
C:\Users\trann>kubect1 scale --replicas=4 statefulset/kafka
statefulset.apps/kafka scaled

C:\Users\trann>_
```

<input type="checkbox"/> kafka	hadoop-ecosystem	4/4	4	23h
<input type="checkbox"/> superset-postgresql	hadoop-ecosystem	1/1	1	26h
<input type="checkbox"/> superset-redis-master	hadoop-ecosystem	1/1	1	26h
<input type="checkbox"/> zookeeper	hadoop-ecosystem	2/2	2	23h

5. Kết quả thực nghiệm

Tính chịu lỗi



5. Kết quả thực nghiệm (Kết quả truy vấn dữ liệu)

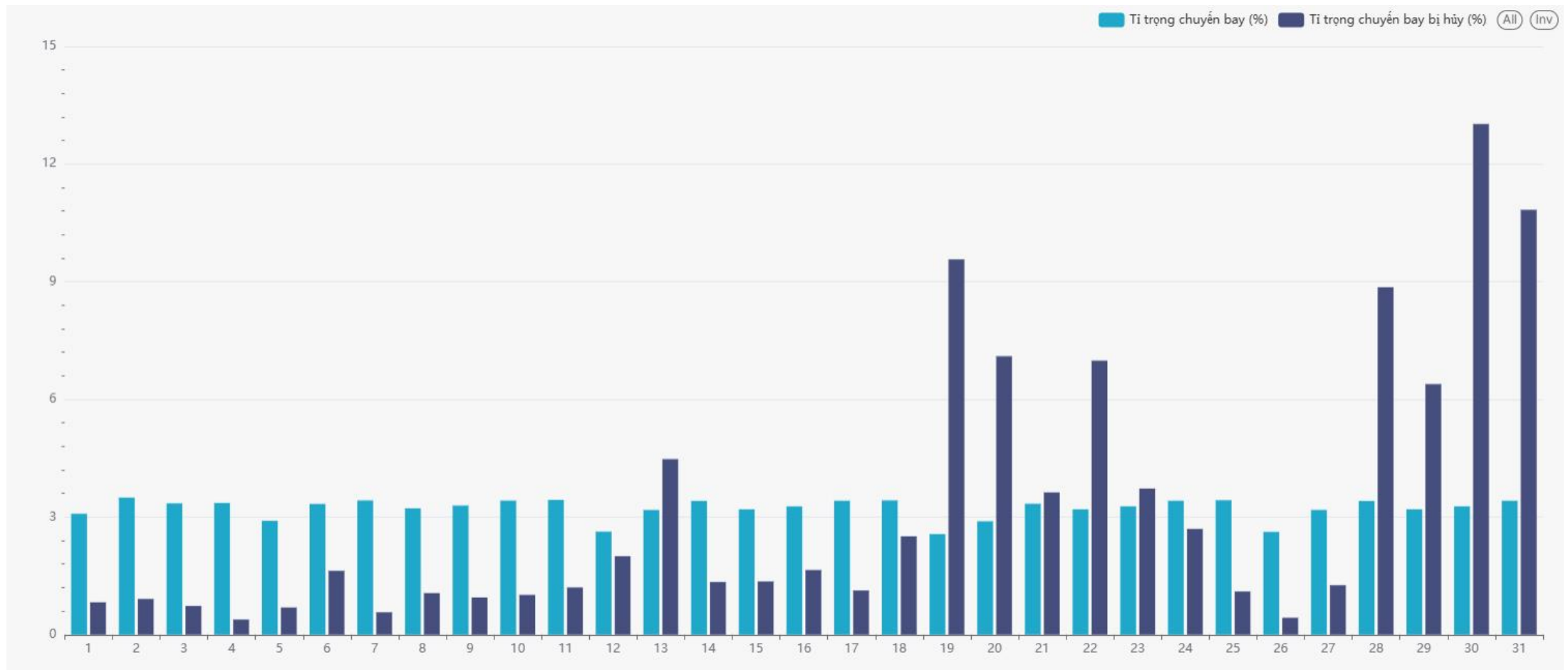
2 hình bên là thông tin đánh giá câu truy vấn tạo bảng "marketing_airline_network_{year}" (phân phối mạng lưới hàng không) trong cả năm 2019.

Thuộc tính	Giá trị
Thời gian đợi	423.00 us
Thời gian phân tích	180.17 ms
Thời gian lên kế hoạch	350.99 ms
Thời gian thực hiện	8.53 s
Tổng số bản ghi	8.09 triệu
Kích thước dữ liệu	2.51 GB

Session		Execution	
User	airflow	Resource Group	global
Principal	airflow	Submission Time	2025-01-05 2:00pm
Source	trino-cli	Completion Time	2025-01-05 2:00pm
Catalog		Elapsed Time	8.71s
Schema		Queued Time	423.00us
Time zone	UTC	Analysis Time	180.17ms
Client Address	172.18.0.10	Planning Time	350.99ms
Client Tags		Execution Time	8.53s
Session Properties			
Resource Estimates			

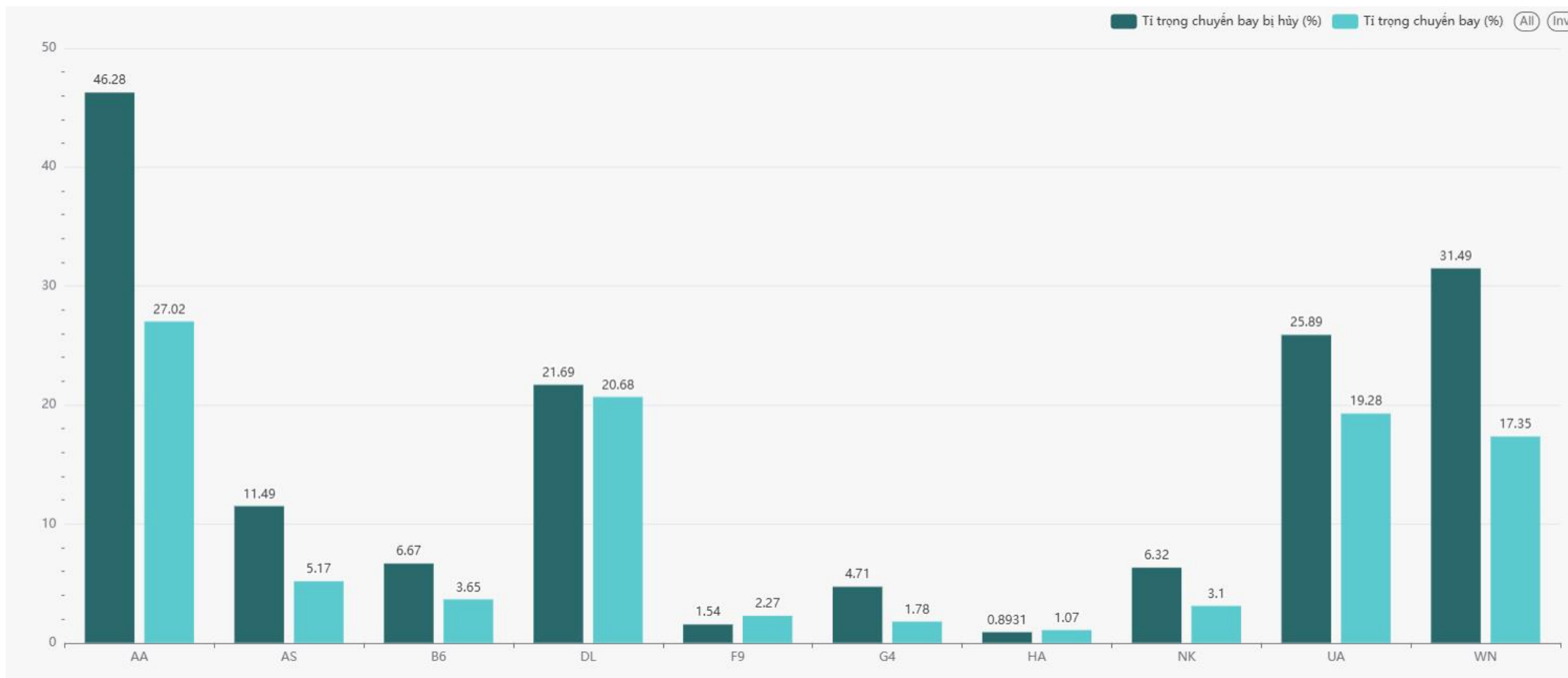
Resource Utilization Summary		Timeline	
CPU Time	51.04s	Parallelism	
Planning CPU Time	197.30ms		5.86
Scheduled Time	4.61m	Scheduled Time/s	
Input Rows	8.09M		31.8
Input Data	170MB	Input Rows/s	
Physical Input Rows	8.09M		929K
Physical Input Data	2.51GB	Input Bytes/s	
Physical Input Read Time	58.75s		19.5MB
Internal Network Rows	1.25K	Physical Input Bytes/s	
Internal Network Data	74.5KB		43.8MB
Peak User Memory	80.3MB	Memory Utilization	
Peak Total Memory	80.3MB		0B
Cumulative User Memory	191MB*seconds		
Output Rows	1.00		
Output Data	9.00B		
Written Rows	120		
Logical Written Data	3.63KB		
Physical Written Data	1.64KB		

5. Kết quả thực nghiệm (Kết quả trực quan hóa dữ liệu)



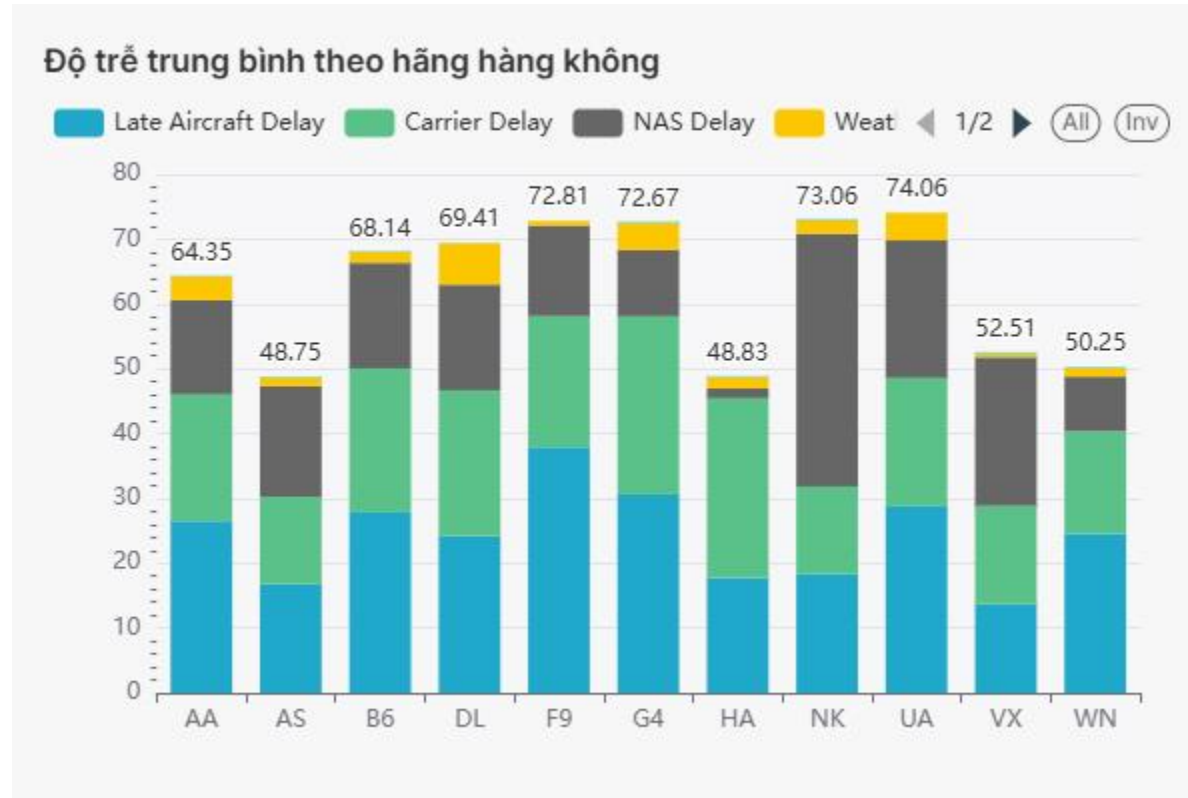
Tỉ trọng chuyến bay và tỉ trọng chuyến bay bị hủy theo ngày trong tháng 1 năm 2019

5. Kết quả thực nghiệm (Kết quả trực quan hóa dữ liệu)



Tỉ trọng chuyến bay và chuyến bay bị hủy theo từng hãng hàng không trong quý 4 năm 2021

5. Kết quả thực nghiệm (Kết quả trực quan hóa dữ liệu)



Độ trễ trung bình theo hãng hàng không
năm 2018

5. Kết quả thực nghiệm (Kết quả trực quan hóa dữ liệu)



Số chuyến bay theo tháng qua các năm



Nội dung

1. Giới thiệu đề tài
2. Mục tiêu và giải pháp
3. Thiết kế hệ thống
4. Triển khai hệ thống
5. Kết quả thực nghiệm
6. Kết luận

6. Kết luận

- Đề án đã xây dựng thành công hệ thống hồ dữ liệu phân tích chuyển bay dựa trên hệ sinh thái Hadoop, đáp ứng yêu cầu lưu trữ và xử lý dữ liệu lớn với chi phí thấp, độ chính xác cao và khả năng tự động hóa.
- Hệ thống hỗ trợ phát dữ liệu trực tuyến, xử lý thời gian thực, mở rộng linh hoạt và trực quan hóa dữ liệu.

6. Kết luận

- Cần triển khai bảo mật dữ liệu, tối ưu hóa tài nguyên hệ thống, phát triển tính năng quản lý chất lượng dữ liệu.
- Tích hợp học máy và trí tuệ nhân tạo để huấn luyện và ứng dụng mô hình theo thời gian thực.
- Mở rộng hệ thống sang hạ tầng đám mây để tận dụng khả năng mở rộng, hiệu suất và chi phí hiệu quả hơn.



HUST

THANK YOU !