

# Big Data Ingestion Tools

- Giới thiệu về Data Ingestion trong hệ thống Big Data.
- Thực hiện Data Ingestion từ nhiều nguồn sử dụng Apache Sqoop và Apache NiFi.
- Thực hành: Cài đặt và sử dụng công cụ Sqoop và Apache NiFi đồng bộ dữ liệu giữa RDBMS và HDFS

Giảng Viên: Nguyễn Chí Thanh





**Nguyễn Chí Thanh**  
Big Data Engineer/ Data Architect  
Blog: <https://karcuta.medium.com>

## ABOUT ME

---

- Trên 5 năm kinh nghiệm trong lĩnh vực Big Data Engineering.
- Tham gia xây dựng và triển khai hệ thống vBI, Viettel Data Lake cho Viettel Telecom.
- Sở hữu chứng chỉ Quốc tế về Hadoop, Spark do Cloudera, Databricks cấp (CCA 175, CRT020).
- Thiết kế phát triển các hệ thống trên nền tảng Hadoop Ecosystem: Hdfs, Spark, Kafka, Hive...

# 1. Data Ingestion

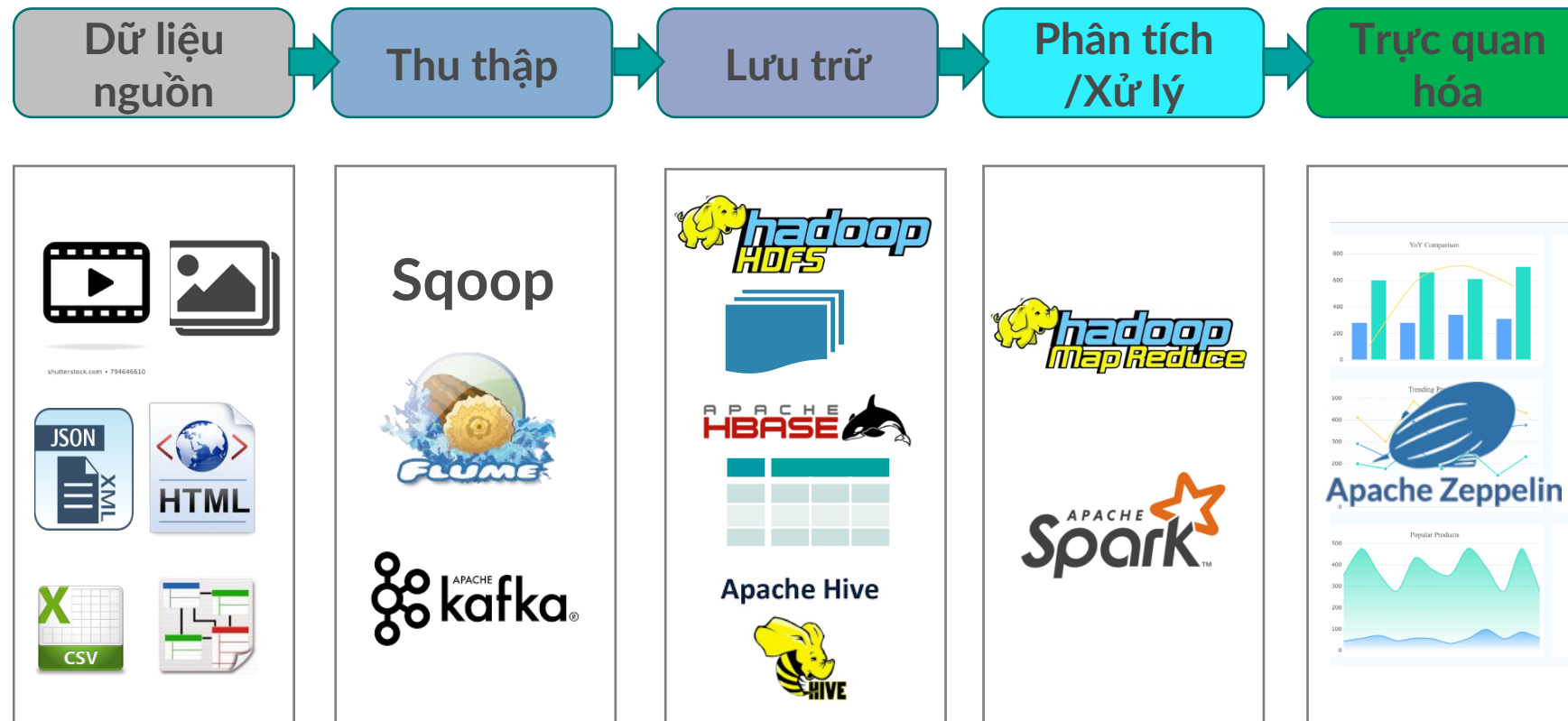


# Các nguồn dữ liệu



- Rất hiếm hệ thống Big Data mà không pull data từ nhiều nguồn khác về.
- Các nguồn dữ liệu rất đa dạng về loại dữ liệu, VD dữ liệu CSV, email, API, bảng trong CSDL.
- Mỗi nguồn dữ liệu đều có các công nghệ khác nhau: FTP, RDBMS, Streaming.

# Phân tích dữ liệu lớn (Big Data Analytics)



# Phân tích dữ liệu lớn (Big Data Analytics)

- Thu thập dữ liệu

- Sqoop: CSV, SQL, MySQL
- Flume
- Kafka
- NiFi

- Lưu trữ dữ liệu

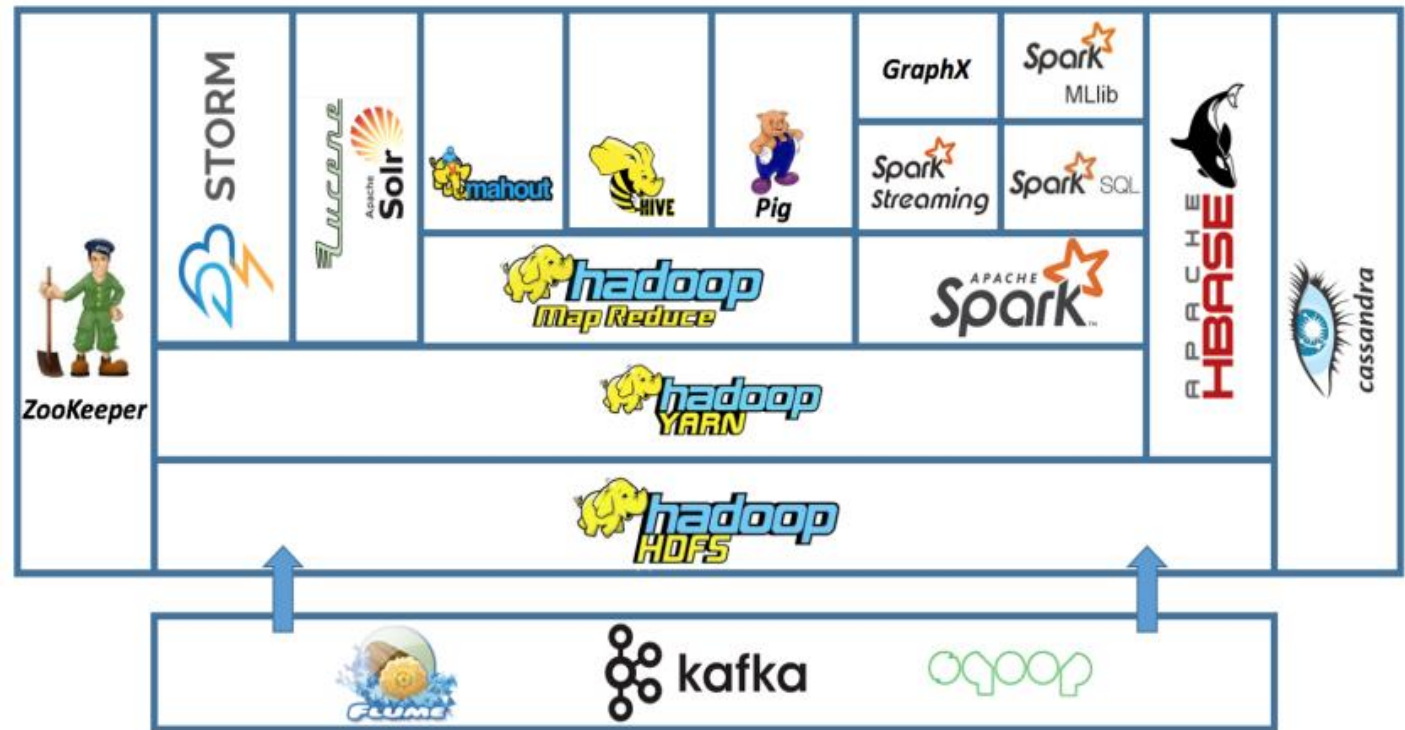
- HDFS
- Hive
- Hbase

- Xử lý dữ liệu

- Hadoop
- Spark

- Trực quan hóa dữ liệu

- Zeppelin



# Data Ingestion

- Data Ingestion là một quá trình mà dữ liệu được di chuyển từ một hoặc nhiều nguồn đến đích nơi dữ liệu có thể được lưu trữ và phân tích thêm.
- Dữ liệu có thể ở các định dạng khác nhau và đến từ nhiều nguồn khác nhau: FPT, RDBMS, Streaming
- Nó cần được làm sạch và chuyển đổi theo cách cho phép bạn phân tích nó cùng với dữ liệu từ các nguồn khác.





# Tích hợp theo lô (Batch Ingestion)



## Tích hợp theo lô

Thường xử lý dữ liệu dạng extract file.



## Lượng lớn dữ liệu

Có thể lên đến hàng triệu/ tỉ bản ghi mỗi lần xử lý.



## Định kỳ

Dạng xử lý này sẽ thường dựa vào việc chạy các job/ tiến trình Tích hợp dữ liệu định kì được đặt lịch trước



## Chi phí vận hành thấp

Không yêu cầu hệ thống vận hành liên tục, phù hợp với doanh nghiệp không có đội ngũ IT chuyên trách





# Real time/ near-real time Ingestion

## Độ trễ thấp

Dữ liệu được tích hợp về hệ thống gần như tức thời hoặc độ trễ rất nhỏ, vài giây cho tới vài phút.



## Vận hành phức tạp

Thông thường các hệ thống real time/ near realtime có yêu cầu cao hơn về phần cứng, phần mềm và đội ngũ nhân sự giám sát, vận hành và phát triển.



## Xu hướng trong tương lai

Ngày nay, nhiều hệ thống yêu cầu việc Ingestion dữ liệu theo thời gian thực để đảm bảo quá trình xử lý được liên tục như hệ thống IoT, smartcity...



## 2. Apache NiFi



# Apache NiFi

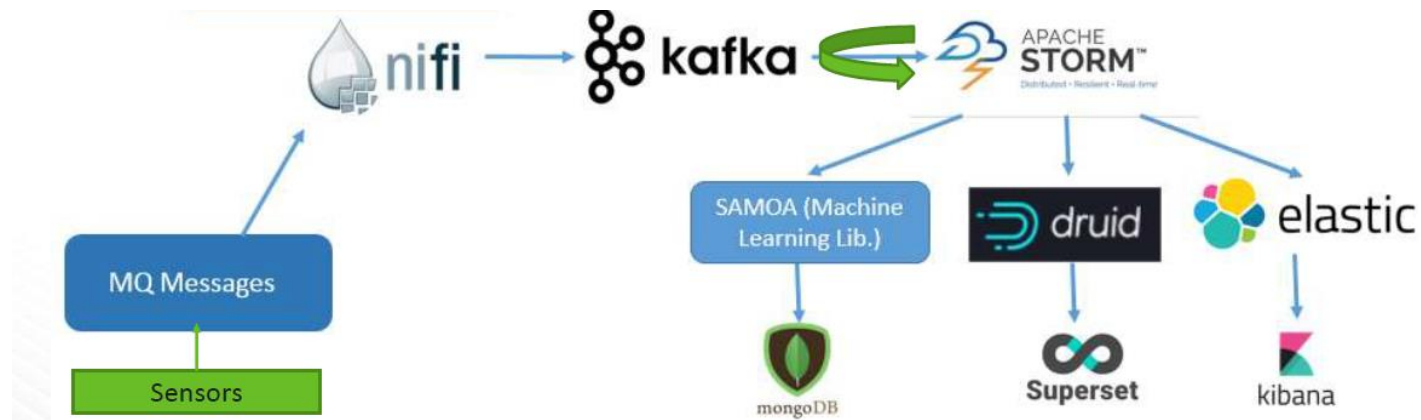
- “Apache NiFi supports powerful and scalable *directed graphs of data* routing, transformation, and system mediation logic.”



# Tính năng của NiFi

## ■ Các chức năng chính

- Tự động hóa luồng dữ liệu giữa các hệ thống
- Ví dụ: JSON -> Database, FTP-> Hadoop, Kafka -> ElasticSearch, etc...
- Giao diện sử dụng kéo thả
- Tập trung vào cấu hình của các khối xử lý (Processor)
- Dễ dàng mở rộng số máy của một cụm
- Đảm bảo không có mất mát dữ liệu
- Data Buffering / Back Pressure / Prioritization Queuing / Latency vs Throughput



# Mục đích sử dụng

## ■ Nên sử dụng NiFi

- Chuyển dữ liệu bảo mật và tin cậy giữa các hệ thống
- Chuyển dữ liệu từ nguồn tới các nền tảng phân tích
- Tiền xử lý dữ liệu
- Thay đổi định dạng dữ liệu
- Trích xuất dữ liệu
- Điều hướng

## ■ Không nên sử dụng NiFi

- Tính toán phân tán
- Xử lý các sự kiện phức tạp
- Thực hiện JOIN, AGGREGATE dữ liệu



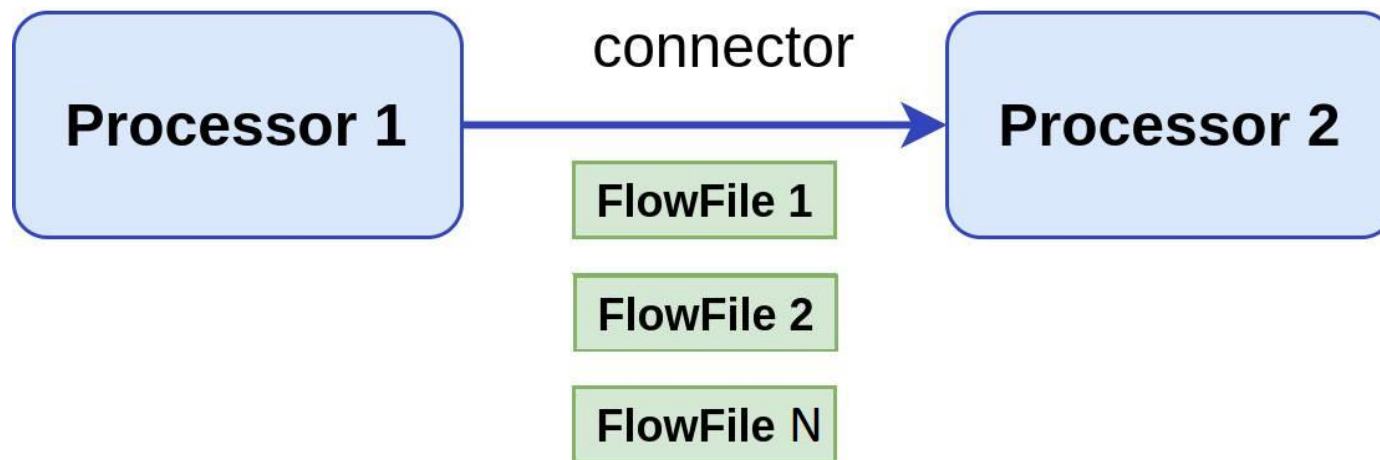
# FlowFile

- Là object đại diện cho dữ liệu đang có trên luồng
  - Gồm có 2 phần
  - Content: dữ liệu thực sự
  - Attributes: cặp key – value liên quan tới dữ liệu
  - Được lưu xuống ổ đĩa ngay sau khi tạo



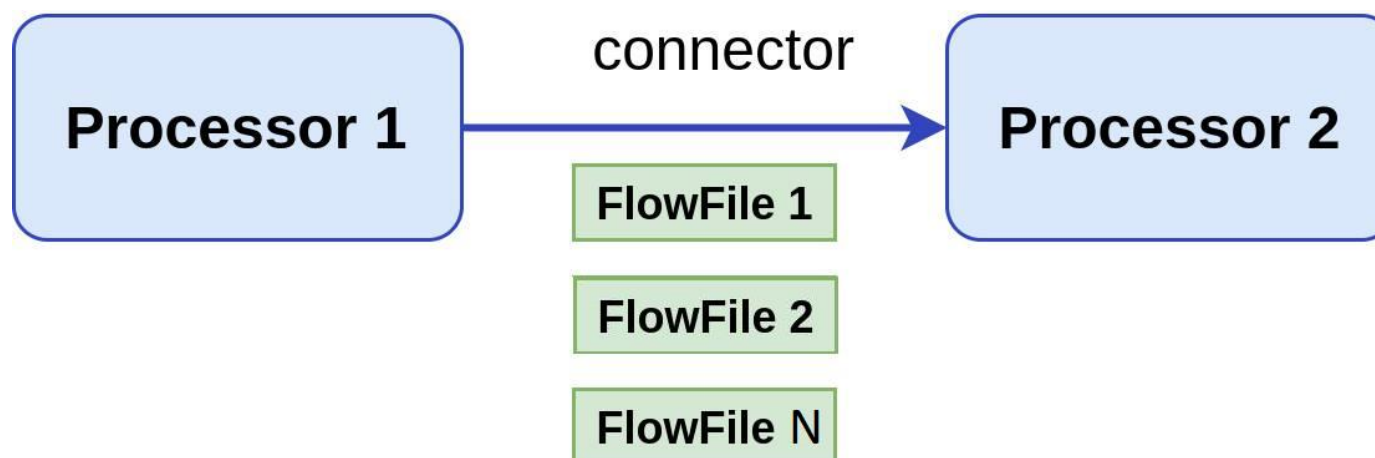
# Processor

- Áp dụng một tập các biến đổi hoặc luật cho FlowFile để tạo ra các FlowFile mới
- Tất cả các Processor đều có thể xử lý được mọi FlowFile đi đến nó
- Chúng chuyển tham chiếu FlowFile cho lẫn nhau để nâng cao quá trình xử lý
- Các processor hoạt động song song trên các thread khác nhau

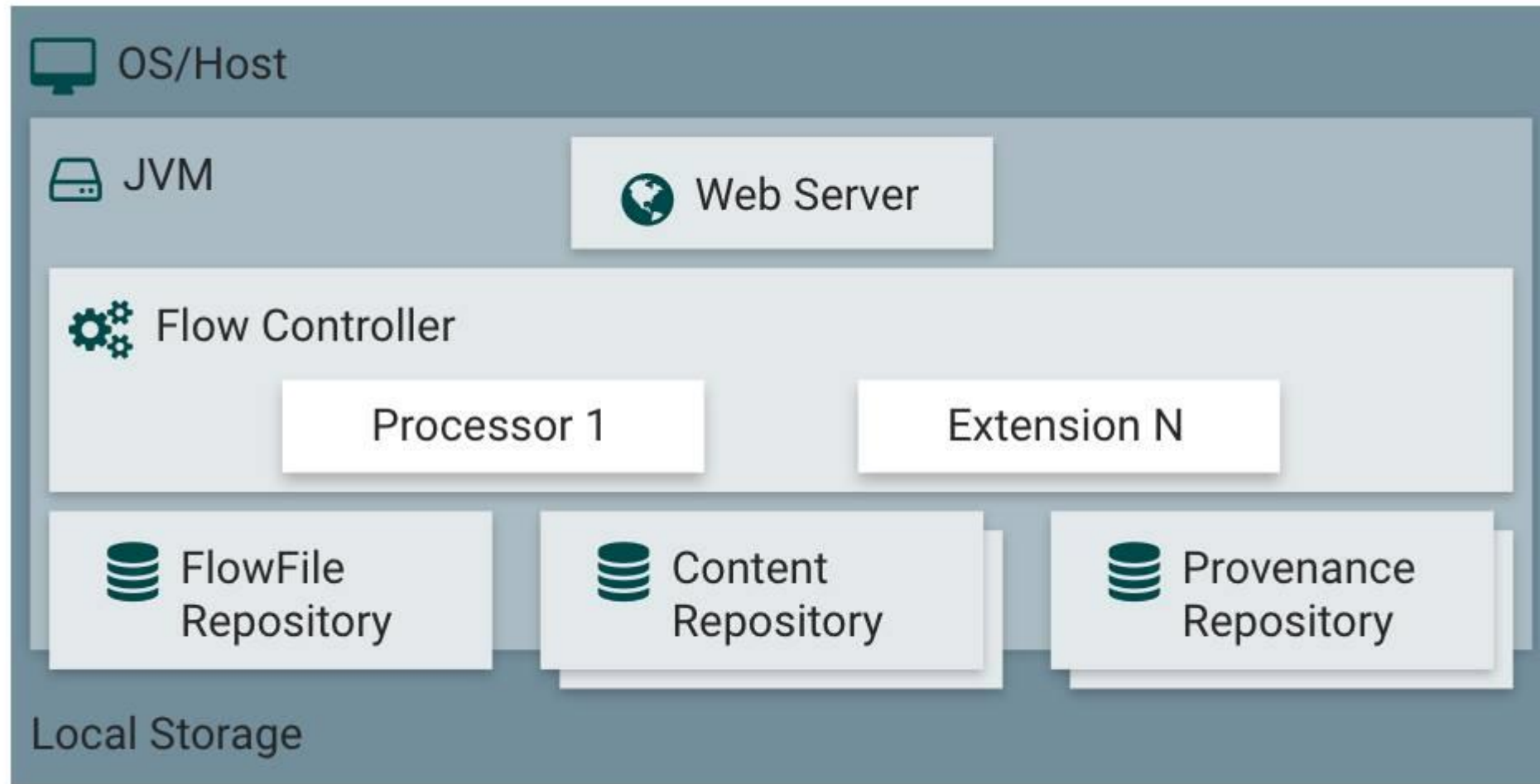


# Connector

- Là hàng đợi của tất cả các FlowFile chưa được xử lý bởi Processor 2
- Định nghĩa các luật về thứ tự ưu tiên xử lý cho các FlowFile
- Có thể đặt ngưỡng backpressure để tránh bị quá tải hệ thống



# NiFi Components



# NiFi Components



Web Server

- Xử lý các request HTTP khi người dùng thao tác với giao diện hoặc thông qua API

Extension N

- Là một lớp các thành phần xây dựng lên luồng dữ liệu trong NiFi bao gồm: Processor, ControllerService, ReportingTask, Prioritizer, ...



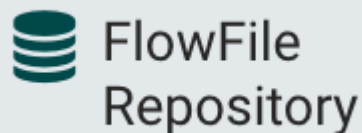
Flow Controller

Extension N

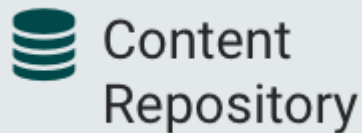
- Là trung tâm điều phối hoạt động và quản lý tài nguyên trong quá trình hoạt động của hệ thống.
- Tạo ra thread cho extension chạy trên đó và quản lý thời gian chạy của extension



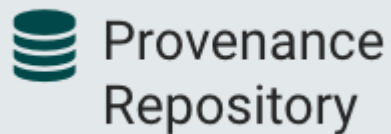
# NiFi Components



- Xử lý các request HTTP khi người dùng thao tác với giao diện hoặc thông qua API.



- Là nơi lưu giữ dữ liệu thực mà các FlowFile đang quản lý.



- Là nơi lưu giữ lại toàn bộ lịch sử xử lý của FlowFile

## Một số nhóm Processor

- **Data Transformation:** ReplaceText, JoltTransformJSON...
- **Routing and Mediation:** RouteOnAttribute, RouteOnContent, ControlRate...
- **Database Access:** ExecuteSQL, ConvertJSONToSQL, PutSQL...
- **Attribute Extraction:** EvaluateJsonPath, ExtractText, UpdateAttribute...
- **System Interaction:** ExecuteProcess ...
- **Data Ingestion:** GetFile, GetFTP, GetHTTP, GetHDFS, ListenUDP, GetKafka...
- **Sending Data:** PutFile, PutFTP, PutKafka, PutEmail...
- **Splitting and Aggregation:** SplitText, SplitJson, SplitXml, MergeContent...
- **HTTP:** GetHTTP, ListenHTTP, PostHTTP...
- **AWS:** FetchS3Object, PutS3Object, PutSNS, GetSQS

## Các loại đầu ra của một Processor

- Các Processor khác nhau sẽ có đầu ra khác nhau. Một số đầu ra cơ bản sau:
  - Success: FlowFile được xử lý thành công
  - Failure: FlowFile ban đầu đi vào Processor và không được xử lý thành công
  - Origin: FlowFile ban đầu đi vào Processor
  - Route: FlowFile được lọc theo điều kiện

3.

## Thực hành – Chuẩn bị dữ liệu



## Chuẩn bị dữ liệu (1)

```
SET PASSWORD FOR 'root'@'localhost' = PASSWORD('ai@acad');
```

```
create database aiacad;
```

```
use aiacad;
```



## Chuẩn bị dữ liệu (2)

```
CREATE TABLE IF NOT EXISTS students(  
    id int,  
    name varchar(255)  
);
```

```
CREATE TABLE IF NOT EXISTS score_sheet(  
    student_id int,  
    cpa float,  
    gpa float  
);
```

## Chuẩn bị dữ liệu (3)

```
insert into students values (1, "Thanh");  
insert into students values (2, "Mai");  
insert into students values (3, "Duc");  
insert into students values (4, "Ha");  
insert into students values (5, "Phu");  
insert into students values (6, "Duy");
```

## Chuẩn bị dữ liệu (4)

```
insert into score_sheet values (1, 3.2, 2.5);  
insert into score_sheet values (2, 2.5, 1.5);  
insert into score_sheet values (4, 3.4, 1.7);  
insert into score_sheet values (6, 1.2, 2.5);
```

## 4. Cài đặt NiFi



# Cài đặt NiFi

- `cd /usr/local`
- `wget https://archive.apache.org/dist/nifi/1.1.1/nifi-1.1.1-bin.tar.gz`
- `tar -xvf nifi-1.1.1-bin.tar.gz`
- `mv nifi-1.1.1/ nifi`
- `cd /usr/local/nifi`
  
- `/bin/nifi.sh start`
- `/bin/nifi.sh status`

**NiFi UI:** `http://master:8080/nifi`

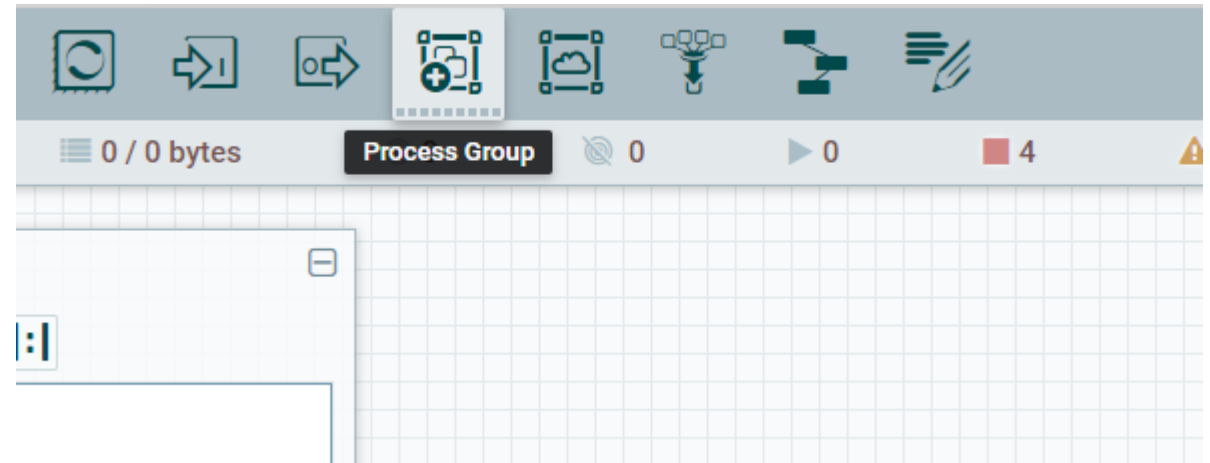


## 5. Thực hành NiFi



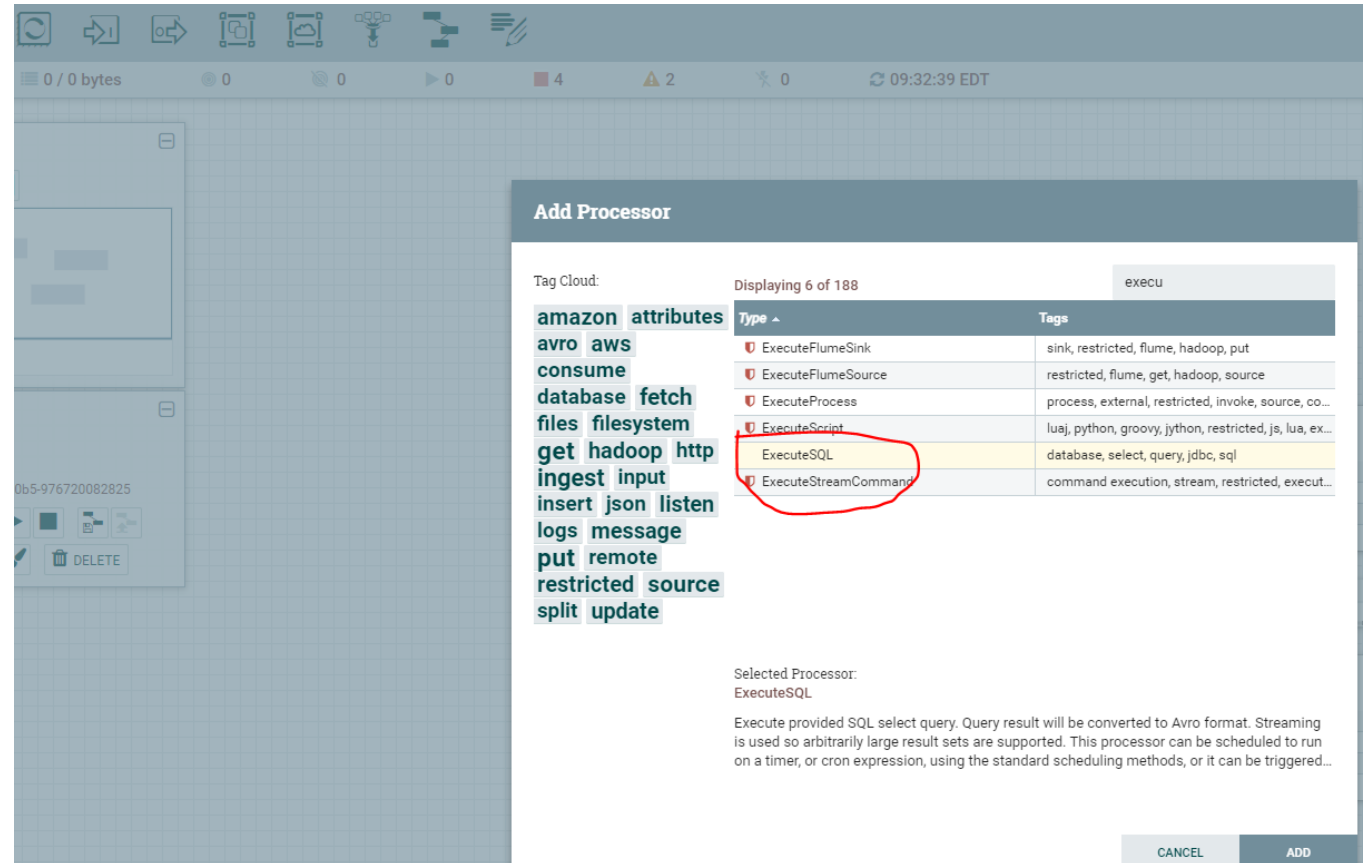
# RDMBS to HDFS (1)

- Tạo process group



## RDMBS to HDFS (2)

- Thêm processor ExecuteSQL



0 / 0 bytes 0 0 0 4 2 0 09:32:39 EDT

**Add Processor**

Tag Cloud: amazon attributes avro aws consume database fetch files filesystem get hadoop http ingest input insert json listen logs message put remote restricted source split update

Displaying 6 of 188

Type ^	Tags
ExecuteFlumeSink	sink, restricted, flume, hadoop, put
ExecuteFlumeSource	restricted, flume, get, hadoop, source
ExecuteProcess	process, external, restricted, invoke, source, co...
ExecuteScript	lua, python, groovy, jython, restricted, js, lua, ex...
ExecuteSQL	database, select, query, jdbc, sql
ExecuteStreamCommand	command execution, stream, restricted, execut...

Selected Processor:  
ExecuteSQL

Execute provided SQL select query. Query result will be converted to Avro format. Streaming is used so arbitrarily large result sets are supported. This processor can be scheduled to run on a timer, or cron expression, using the standard scheduling methods, or it can be triggered...

CANCEL ADD

## RDMBS to HDFS (3)

- Vào tab properties, cấu hình DB Connection services pool
- Bấm vào mũi tên để sang phần config

### Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field



Property		Value	
Database Connection Pooling Service	?	DBCConnectionPool	→
SQL select query	?	select * from students	
Max Wait Time	?	0 seconds	
Normalize Table/Column Names	?	false	

## RDMBS to HDFS (4)

### ■ Thêm Controller services:

- jdbc:mysql://localhost:3306/aiacad
- com.mysql.jdbc.Driver
- /usr/local/nifi/lib/mysql-connector-java-5.1.49.jar
- root
- ai@acad

### Configure Controller Service

SETTINGS

PROPERTIES

COMMENTS

Required field +

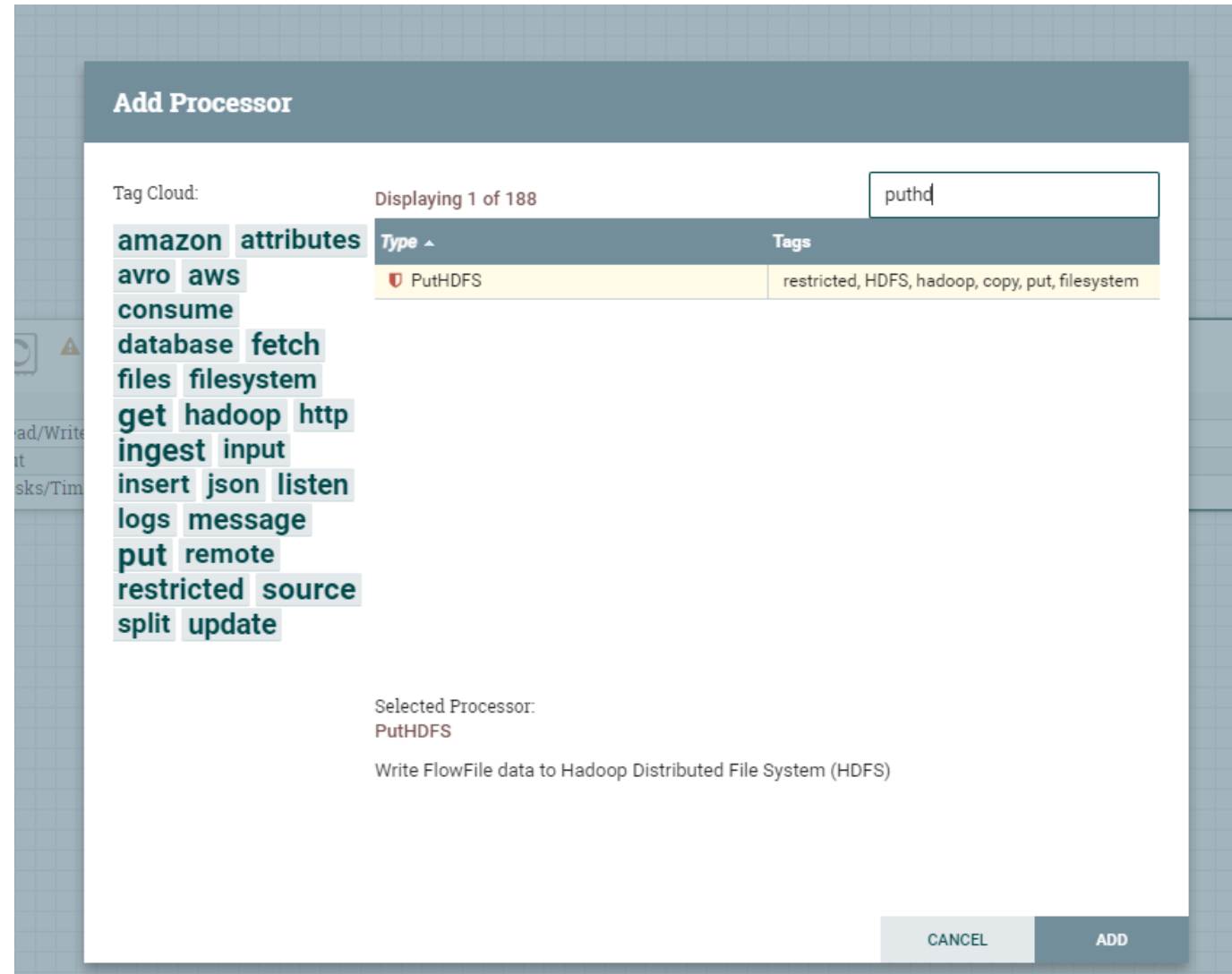
Property		Value
Database Connection URL	?	jdbc:mysql://localhost:3306/aiacad
Database Driver Class Name	?	com.mysql.jdbc.Driver
Database Driver Location(s)	?	/home/hadoop/sqoop-1.4.7.bin__hadoop-2.6.0/lib...
Database User	?	root
Password	?	Sensitive value set
Max Wait Time	?	500 millis
Max Total Connections	?	8
Validation query	?	No value set

CANCEL

APPLY

## RDMBS to HDFS (5)

- Thêm processor PutHDFS



# RDMBS to HDFS (6)

- Sửa properties của PutHDFS Processor
  - /usr/local/hadoop/etc/hadoop/core-site.xml,  
/usr/local/hadoop/etc/hadoop/hdfs-site.xml
  - /aiacad/nifi/students

The screenshot shows the 'Configure Processor' dialog for the PutHDFS processor in NiFi. The 'PROPERTIES' tab is selected, displaying a table of configuration properties. The 'Directory' property is highlighted in yellow and set to '/aiacad/nifi/students'. Other properties include Hadoop Configuration Resources, Kerberos Principal, Kerberos Keytab, Kerberos Rlogin Period, Additional Classpath Resources, Conflict Resolution Strategy, Block Size, IO Buffer Size, Replication, Permissions umask, Remote Owner, Remote Group, and Compression codec.

Property	Value
Hadoop Configuration Resources	/usr/local/hadoop/etc/hadoop/core-site.xml, /u...
Kerberos Principal	No value set
Kerberos Keytab	No value set
Kerberos Rlogin Period	4 hours
Additional Classpath Resources	No value set
Directory	/aiacad/nifi/students
Conflict Resolution Strategy	fail
Block Size	No value set
IO Buffer Size	No value set
Replication	No value set
Permissions umask	No value set
Remote Owner	No value set
Remote Group	No value set
Compression codec	NONE



# RDMBS to HDFS (7)

- Sửa properties của PutHDFS Processor
  - /usr/local/hadoop/etc/hadoop/core-site.xml,  
/usr/local/hadoop/etc/hadoop/hdfs-site.xml
  - /aiacad/nifi/students

**Configure Processor**

SETTINGS SCHEDULING **PROPERTIES** COMMENTS

Required field +

Property		Value
Hadoop Configuration Resources	?	/usr/local/hadoop/etc/hadoop/core-site.xml, /u...
Kerberos Principal	?	No value set
Kerberos Keytab	?	No value set
Kerberos Rlogin Period	?	4 hours
Additional Classpath Resources	?	No value set
<b>Directory</b>	?	<b>/aiacad/nifi/students</b>
<b>Conflict Resolution Strategy</b>	?	<b>fail</b>
Block Size	?	No value set
IO Buffer Size	?	No value set
Replication	?	No value set
Permissions umask	?	No value set
Remote Owner	?	No value set
Remote Group	?	No value set
Compression codec	?	NONE

CANCEL APPLY

## RDMBS to HDFS (8)

- Ở Processor *Execute SQL*, ở tab Setting, chọn failure như hình

### Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Name

ExecuteSQL

☒ Enabled

Automatically Terminate Relationships ?

☒ failure

SQL query execution failed. Incoming FlowFile will be penalized and routed to this relationship

☐ success

Successfully created FlowFile from SQL query result set.

Id

96b926be-0178-1000-d210-1a223e24948a

Type

ExecuteSQL

Penalty Duration ?

30 sec

Yield Duration ?

1 sec

Bulletin Level ?

WARN

## RDMBS to HDFS (9)

- Ở Processor *PutHDFS*, ở tab Setting, chọn failure và success như hình

### Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Name PutHDFS	<input checked="" type="checkbox"/> Enabled	Automatically Terminate Relationships ?	
Id 96bd79fd-0178-1000-3acd-7919f879e40b		<input checked="" type="checkbox"/> failure Files that could not be written to HDFS for some reason are transferred to this relationship	
Type PutHDFS		<input checked="" type="checkbox"/> success Files that have been successfully written to HDFS are transferred to this relationship	
Penalty Duration ? 30 sec	Yield Duration ? 1 sec		
Bulletin Level ? WARN			

## RDMBS to HDFS (10)

- Kết nối 2 processor. Phần relationships chọn success.

**Create Connection**

DETAILS

SETTINGS

From Processor

ExecuteSQL

ExecuteSQL

Within Group

Demo\_AIAcad

For Relationships

☐ failure

☒ success

To Processor

PutHDFS

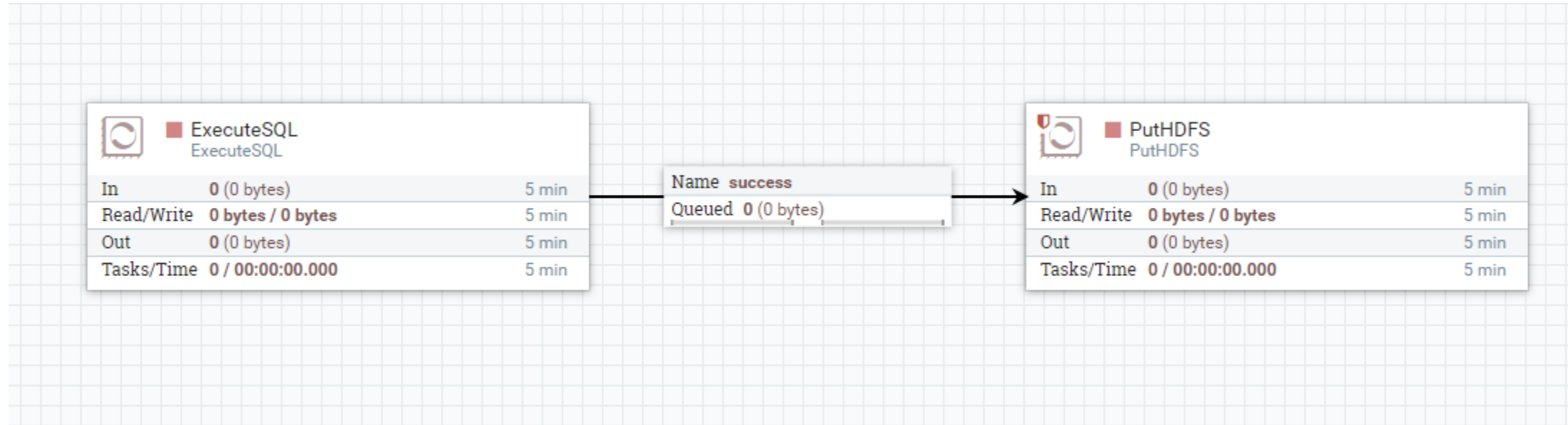
PutHDFS

Within Group

Demo\_AIAcad

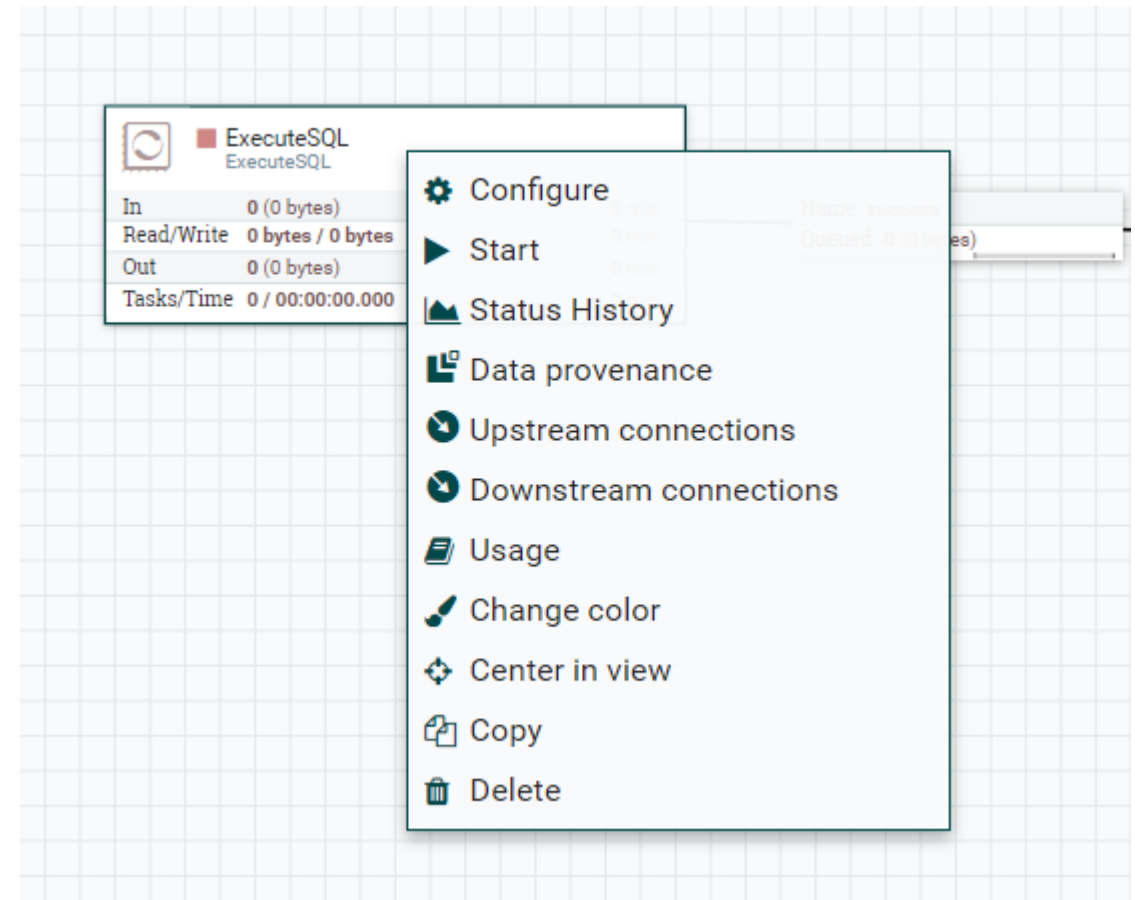
## RDMBS to HDFS (11)

- Kết nối 2 processor.



## RDMBS to HDFS (12)

- Kết nối 2 processor.
- Start từng processor



# RDMBS to HDFS (13)

- Check kết quả ở thư mục:  
/aiacad/nifi/students

/aiacad/nifi/students

Go!

Show

25

▼

entries

Search:

<input type="checkbox"/>	<div><div>⌵⌵</div>Permission</div>	<div><div>⌵⌵</div>Owner</div>	<div><div>⌵⌵</div>Group</div>	<div><div>⌵⌵</div>Size</div>	<div><div>⌵⌵</div>Last Modified</div>	<div><div>⌵⌵</div>Replication</div>	<div><div>⌵⌵</div>Block Size</div>	<div><div>⌵⌵</div>Name</div>	<div><div>⌵⌵</div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749118361996</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749135230523</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749142325181</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749161996730</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749165604029</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749169435837</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749171646116</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749173449115</a>	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	287 B	Apr 03 15:32	<a href="#">1</a>	128 MB	<a href="#">2749216686827</a>	<div><div></div></div>