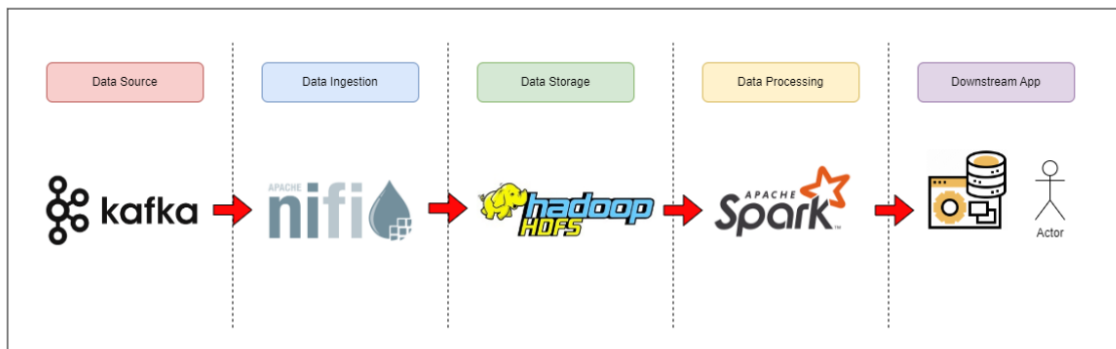


AssignmentDE

- Kiến trúc hệ thống



- Kafka gồm 2 broker và được quản lý bởi zookeeper
- Nifi bao gồm NiFi registry và Nifi được quản lý bởi zookeeper
- HDFS gồm namenode, 2 datanode cùng với nodemanager và resourcemanager phục vụ cho YARN
- Spark gồm 1 master node, 2 worker node và 1 spark-submit

▼ Cách thức triển khai, thực hiện các bước

- Cấu hình các container như trong docker-compose.yml
- Các dependency cần thiết có trong pom.xml

▼ Set up phần Nifi

Ban đầu chúng ta chưa mount thư mục conf vào Nifi

```
- ./nifi/provenance_repository:/opt/nifi/nifi-current/provenance_repository
- ./nifi/state:/opt/nifi/nifi-current/state
- ./nifi/logs:/opt/nifi/nifi-current/logs
# uncomment the next line after copying the /conf directory from the container to your local directory
#- ./nifi/conf:/opt/nifi/nifi-current/conf
networks:
  - persistent_network
```

Sau khi docker compose up thì các thư mục chưa tồn tại sẽ được tạo ra trên local và chúng ta phải tìm ID của container Nifi để copy thư mục conf từ Nifi về local

```
# to get the container ID of NiFi's docker container
```

```
> docker ps
```

```
# the result will look like this (shortened to fit this article)
```

CONTAINER ID	IMAGE	COMMAND
7554d9c68c8f	apache/nifi:1.14.0	...
8af04cd37e06	apache/nifi-registry:1.15.0	...
a2dacb43ed23	bitnami/zookeeper:3.7.0	...

```
# copy the directory from the docker container to the local machine
```

```
> docker cp 7554d9c68c8f:/opt/nifi/nifi-current/conf ./nifi/
```

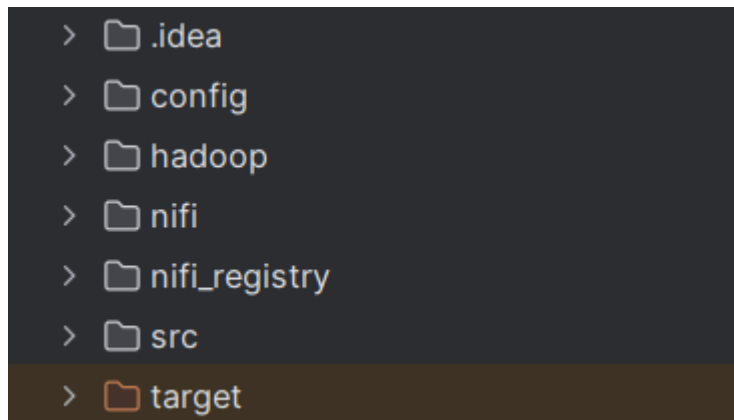
Copy thành công thì lần compose sau chúng ta sẽ bỏ comment câu lệnh

```
- ./nifi/conf:/opt/nifi/nifi-current/conf
```

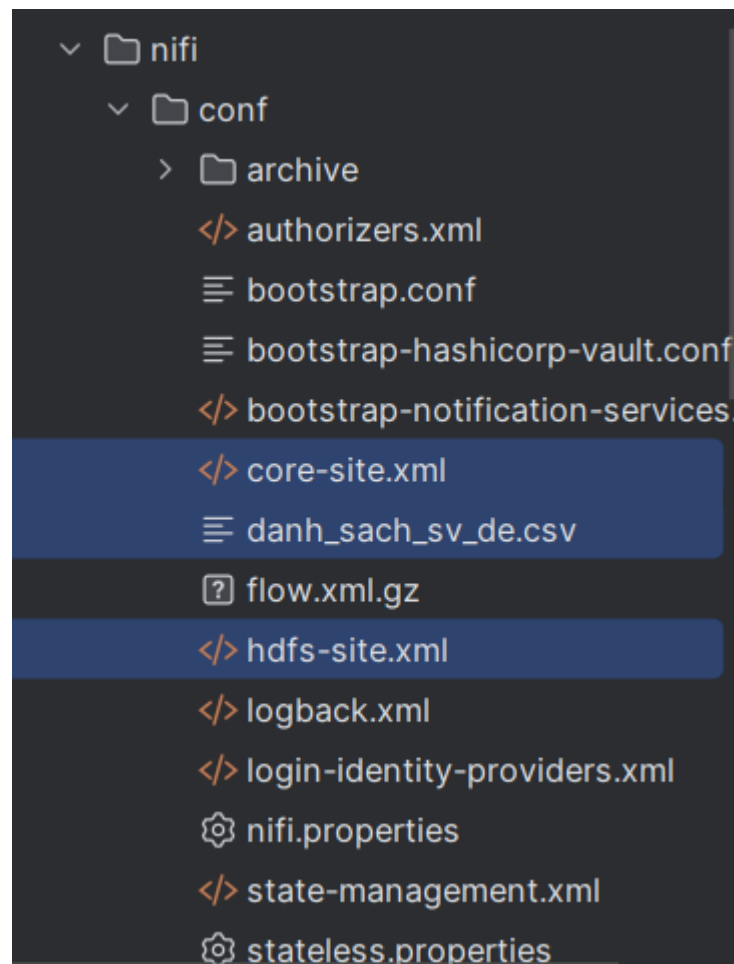
Chúng ta sẽ mount thư mục các file cấu hình của hdfs ra local

```
docker exec -it d4d7897d4c4e9d21fa3279cb0b24f36464d9fcb4fdc6cc3da603b5df0cd3ff3f /bin/sh
sh-4.2$ ls
LICENSE-binary  NOTICE-binary  README.txt  etc  lib  licenses-binary  share
LICENSE.txt     NOTICE.txt     bin         include  libexec  sbin
sh-4.2$ cd etc
sh-4.2$ ls
hadoop
sh-4.2$ cd hadoop
sh-4.2$ ls
capacity-scheduler.xml  hadoop-user-functions.sh.example  kms-site.xml  user_ec_policies.xml.template
capacity-scheduler.xml.raw  hdfs-rbf-site.xml  log4j.properties  workers
configuration.xsl  hdfs-site.xml  mapred-env.cmd  yarn-env.cmd
container-executor.cfg  hdfs-site.xml.raw  mapred-env.sh  yarn-env.sh
core-site.xml  https-env.sh  mapred-queues.xml.template  yarn-site.xml
core-site.xml.raw  https-log4j.properties  mapred-site.xml  yarn-site.xml.raw
hadoop-env.cmd  https-site.xml  mapred-site.xml.raw  yarnservice-log4j.properties
hadoop-env.sh  kms-acls.xml  shellprofile.d
hadoop-metrics2.properties  kms-env.sh  ssl-client.xml.example
hadoop-policy.xml  kms-log4j.properties  ssl-server.xml.example
sh-4.2$ pwd
sh: pwd: command not found
sh-4.2$ pwd
/opt/hadoop/etc/hadoop
sh-4.2$
```

```
PS C:\Users\Administrator\Desktop\HUST\BigData\Project> docker cp d4d7897d4c4e9d21fa3279cb0b24f36464d9fcb4fdc6cc3da603b5df0cd3ff3f:/opt/hadoop/etc/hadoop ./
Successfully copied 120kB to C:\Users\Administrator\Desktop\HUST\BigData\Project\.
```



Copy 2 file core-site.xml và hdfs-site.xml tới ./nifi/conf để phục vụ cho việc truyền file từ Nifi vào HDFS. Copy file danh_sach_sv_de.csv tới ./nifi/conf để phục vụ truyền file sau này.



Đến đây việc chuẩn bị cho Nifi đã hoàn thành.

▼ Producer của Kafka

```

public static void main(String[] args) {
    Properties props = new Properties();
    props.put(ProducerConfig.BOOTSTRAP_SERVERS_CONFIG, BOOTSTRAP_SERVERS);
    props.put(ProducerConfig.KEY_SERIALIZER_CLASS_CONFIG, "org.apache.kafka.common.serialization.StringSerializer");
    props.put(ProducerConfig.VALUE_SERIALIZER_CLASS_CONFIG, "org.apache.kafka.common.serialization.StringSerializer");

    ObjectMapper objectMapper = new ObjectMapper();

    try (KafkaProducer<String, String> producer = new KafkaProducer<>(props);
        BufferedReader reader = new BufferedReader(new FileReader(logPath))) {
        String line;
        int recordCount = 0;

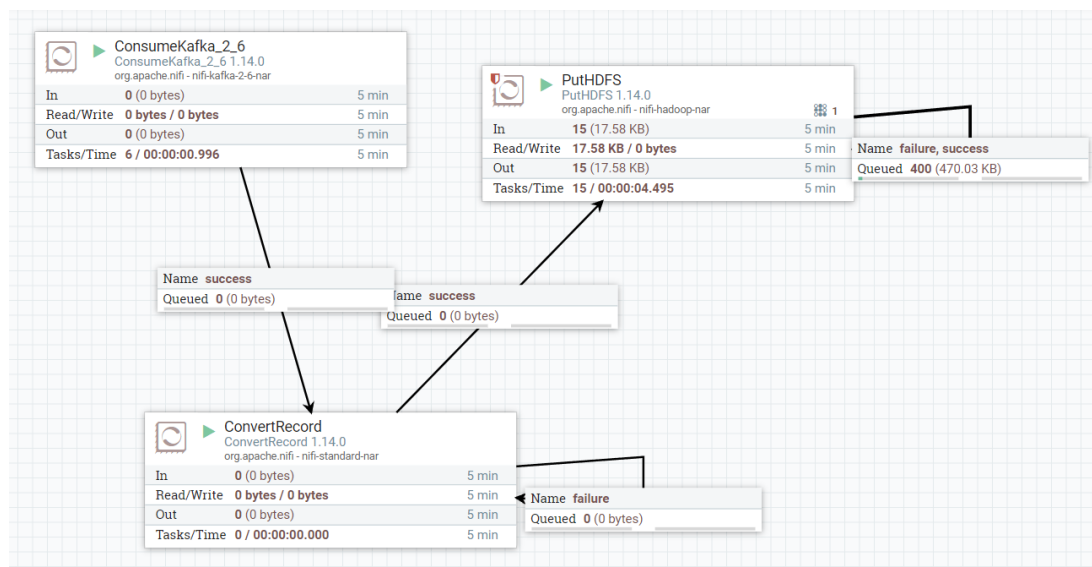
        while ((line = reader.readLine()) != null) {
            // Process the CSV line and create a JSON string
            String[] fields = line.split(regex: ",");
            Map<String, Object> jsonMap = new HashMap<>();
            jsonMap.put(k: "student_code", Integer.parseInt(fields[0]));
            jsonMap.put(k: "activity", fields[1]);
            jsonMap.put(k: "numberOfFile", Integer.parseInt(fields[2]));
            jsonMap.put(k: "timestamp", fields[3]);

            String jsonString = objectMapper.writeValueAsString(jsonMap);
            String key = getKey(fields); // Control partitions of records
            ProducerRecord<String, String> record = new ProducerRecord<>(TOPIC_NAME, key, jsonString);

```

Sau khi broker Kafka hoạt động ổn định chạy Producer sẽ đẩy dữ liệu lên Kafka broker

▼ Triển khai phần Nifi để nhận dữ liệu từ kafka, convert record sang parquet và truyền file parquet tới HDFS để lưu trữ ở đường dẫn `"/raw_zone/fact/activity"`.



- Cấu hình ConsumeKafka

Processor Details

Running

STOP & CONFIGURE

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value
Kafka Brokers	broker01:9093, broker02:9093
Security Protocol	PLAINTEXT
SASL Mechanism	GSSAPI
Kerberos Service Name	No value set
Kerberos Credentials Service	No value set
Kerberos Principal	No value set
Kerberos Keytab	No value set
Username	No value set
Password	No value set
Token Auth	false
SSL Context Service	No value set
Topic Name(s)	vdt2024
Topic Name Format	

OK

Processor Details

Running

STOP & CONFIGURE

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value
Topic Name(s)	vdt2024
Topic Name Format	names
Honor Transactions	true
Group ID	vdt_orders
Offset Reset	latest
Key Attribute Encoding	UTF-8 Encoded
Message Demarcator	No value set
Separate By Key	false
Message Header Encoding	UTF-8
Headers to Add as Attributes (Regex)	No value set
Max Poll Records	10000
Max Uncommitted Time	1 secs
Commitment Timeout	60 secs

OK

- Cấu hình PutHDFS

Processor Details

Running (1)

STOP & CONFIGURE

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value	
Hadoop Configuration Resources	/opt/nifi/nifi-current/conf/hdfs-site.xml, /opt/nifi/nifi-curre...	
Kerberos Credentials Service	No value set	
Kerberos Principal	No value set	
Kerberos Keytab	No value set	
Kerberos Password	No value set	
Kerberos Relogin Period	4 hours	
Additional Classpath Resources	No value set	
Directory	/raw_zone/fact/activity	
Conflict Resolution Strategy	append	
Block Size	No value set	
IO Buffer Size	No value set	
Replication	No value set	

OK

- Cấu hình ConvertRecord

Processor Details

Running

STOP & CONFIGURE

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value	
Record Reader	JsonTreeReader	→
Record Writer	ParquetRecordSetWriter	→
Include Zero Record FlowFiles	true	

OK

Kết quả sau khi thực hiện

Browse Directory

/raw_zone/fact/activity

Go!

Show

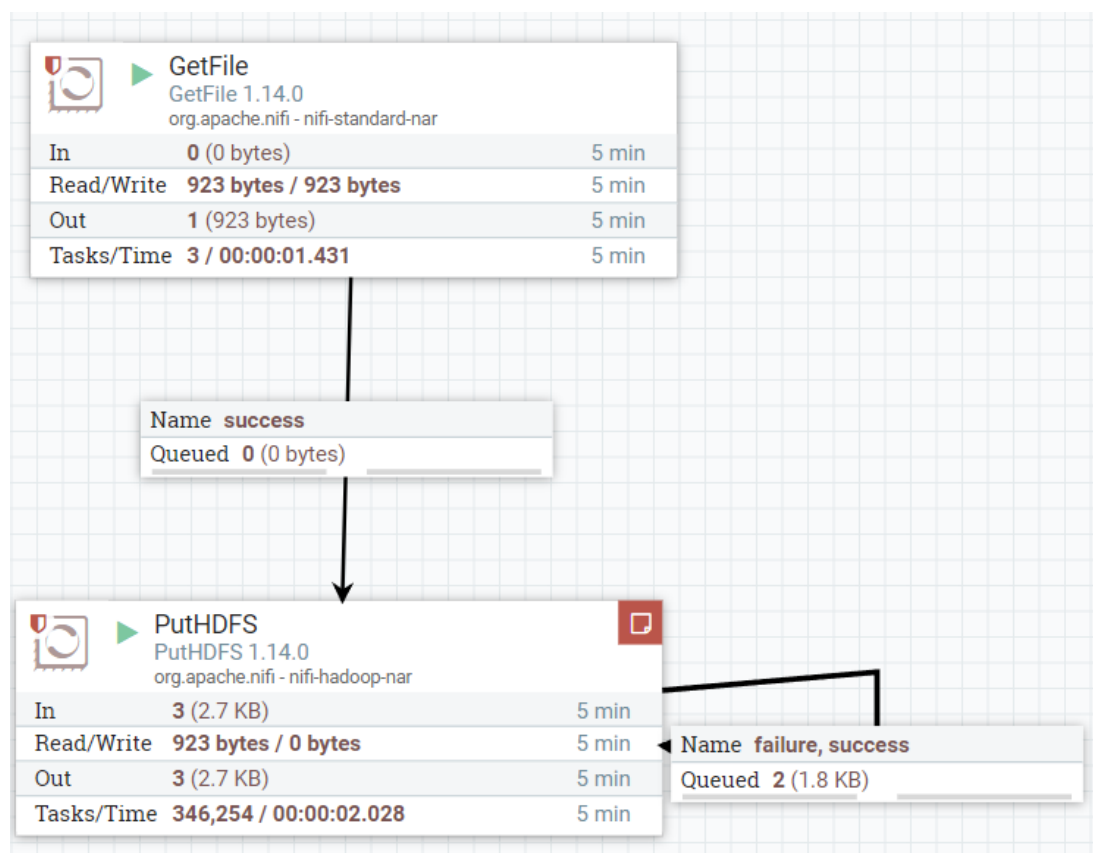
25

entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.19 KB		Jun 12 15:39		1		128 MB		.eb64f96a-4681-4b64-b4eb-960cbc676527	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.19 KB		Jun 12 15:38		1		128 MB		022ee80a-7f2a-4595-9d5b-0bab0f47d3bb	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.17 KB		Jun 12 15:39		1		128 MB		0c01f596-8169-40ed-9101-68e4bbf518b4	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.19 KB		Jun 12 15:38		1		128 MB		1215a35e-d32d-4b2a-864f-5997a31961b3	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.17 KB		Jun 12 15:38		1		128 MB		137b35ab-fd13-41ef-8dfc-e88480ab5818	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.17 KB		Jun 12 15:39		1		128 MB		14b1fbee-ca76-4962-99e2-ea872c69e3ec	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.17 KB		Jun 12 15:38		1		128 MB		16234a96-51d3-48db-9be8-997624c671cb	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.19 KB		Jun 12 15:38		1		128 MB		16e9c726-f493-4da1-a6b2-6da88b353556	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.17 KB		Jun 12 15:38		1		128 MB		17cf3afc-cc05-4d20-a88f-70900d3ccafb	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.19 KB		Jun 12 15:38		1		128 MB		1ab22457-3915-414a-a99f-33faadd0a180	
<input type="checkbox"/>		-rw-r--r--		nifi		supergroup		1.19 KB		Jun 12 15:39		1		128 MB		1c19210d-f4fd-4ecc-850e-db2b80207fd7	

▼ Triển khai Nifi truyền file “danh_sach_sv_de.csv” đến HDFS



Cấu hình PutHDFS như trên.

Cấu hình GetFile

Configure Processor

Stopped

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value
Input Directory	/opt/nifi/nifi-current/conf
File Filter	danh_sach_sv_de.csv
Path Filter	No value set
Batch Size	10
Keep Source File	false
Recurse Subdirectories	true
Polling Interval	0 sec
Ignore Hidden Files	true
Minimum File Age	0 sec
Maximum File Age	No value set
Minimum File Size	0 B
Maximum File Size	No value set

CANCEL

APPLY

Kết quả sau khi thực hiện

Browse Directory

/input

Go!

Show

25

entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	nifi	supergroup	923 B	Jun 12 15:37	1	128 MB	danh_sach_sv_de.csv	

Showing 1 to 1 of 1 entries

Previous

1

Next

Hadoop, 2023.

▼ Apache Spark

Chương trình để đọc danh_sach_sv_de.csv chuyển thành dataframe và đọc lần lượt các file parquet để tạo thành dataframe. Từ đó xử lí dữ liệu và xuất output vào HDFS.


```

SparkSession spark = SparkSession.builder()
    .appName("Read Parquet from HDFS and Aggregate Activities")
    .config("spark.master", "local")
    .getOrCreate();

String studentFilePath = "hdfs://namenode:8020/input/danh_sach_sv_de.csv";

StructType schema = new StructType(new StructField[]{
    DataTypes.createStructField( name: "student_code", DataTypes.IntegerType, nullable: false),
    DataTypes.createStructField( name: "student_name", DataTypes.StringType, nullable: false)
});

Dataset<Row> studentDf = spark.read()
    .format( source: "csv")
    .schema(schema)
    .option("header", "false")
    .load(studentFilePath);

String hdfsPath = "hdfs://namenode:8020/raw_zone/fact/activity";
Dataset<Row> df = spark.read().parquet(hdfsPath);

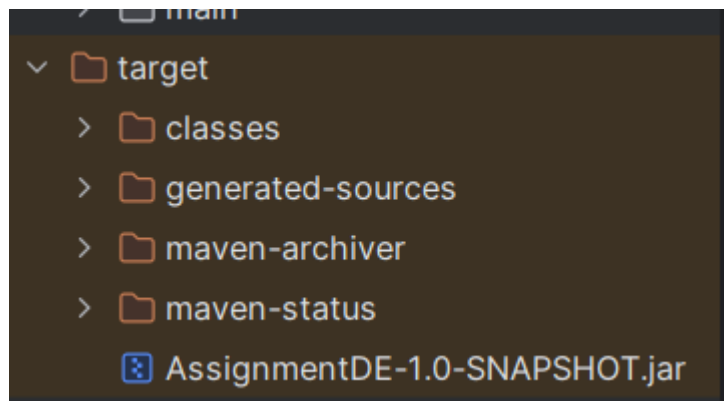
Dataset<Row> aggregatedDF = df.groupBy( col1: "student_code", ...cols: "activity", "timestamp")
    .agg(sum( columnName: "numberOfFile").as( alias: "totalFile"));

Dataset<Row> outputDF = aggregatedDF.join(studentDf, usingColumn: "student_code")
    .select( col: "timestamp", ...cols: "student_code", "student_name", "activity", "totalFile")
    .orderBy( sortCol: "student_code", ...sortCols: "activity");

String outputHdfsPath = "hdfs://namenode:8020/output/aggregated_activities";
outputDF.write()
    .mode( saveMode: "overwrite")
    .option("header", "false")

```

Sau đó chạy lệnh mvn package trong terminal để xuất ra file .jar



Cấu hình spark-submit để khi khởi động sẽ chạy chương trình thực thi quá trình chuyển đổi và đưa ra kết quả

```

spark-submit-job:
  image: bitnami/spark:latest
  container_name: spark-submit-job
  depends_on:
    - spark-master
    - spark-worker-1
    - spark-worker-2
  command: bash -c "sleep 30 && spark-submit --class org.spark.ParquetReader --master spark://spark-master:7077 /app/AssignmentDE-1.0-SNAPSHOT.jar"
  volumes:
    - ./target/AssignmentDE-1.0-SNAPSHOT.jar:/app/AssignmentDE-1.0-SNAPSHOT.jar
  networks:
    - my_persistent_network

```

Do không phải spark streaming nên phải đợi Nifi truyền hết dữ liệu vào HDFS thì mới khởi động container spark-submit để xử lí.

Kết quả sau khi xử lí được lưu vào HDFS

Browse Directory

/output/aggregated_activities

Go!

Show

25

entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	spark	supergroup	0 B	Jun 12 15:41	3	128 MB	._SUCCESS	
<input type="checkbox"/>	-rw-r--r--	spark	supergroup	11.97 KB	Jun 12 15:41	3	128 MB	part-00000-efcbec96-5522-4a0f-852b-65404a81f32b-c000.csv	

Showing 1 to 2 of 2 entries

Previous

1

Next

Hadoop. 2023.

Hadoop, 2023.

Copy file .csv về container

```

docker exec -it 76893e740fa83821571483bb56da8c3259209b1af46384f38cdec63e8688ba94 /bin/sh
sh-4.2$ ls
LICENSE-binary  NOTICE-binary  README.txt  etc  lib  licenses-binary  share
LICENSE.txt     NOTICE.txt     bin         include  libexec  sbin
sh-4.2$ hdfs dfs -get /output/aggregated_activities/part-00000-efcbec96-5522-4a0f-852b-65404a81f32b-c000.csv ./
sh-4.2$ ls
LICENSE-binary  NOTICE.txt  etc  libexec  sbin
LICENSE.txt     README.txt  include  licenses-binary  share
NOTICE-binary   bin         lib      part-00000-efcbec96-5522-4a0f-852b-65404a81f32b-c000.csv
sh-4.2$

```

Copy file từ container về local

```

PS C:\Users\Administrator\Desktop\VDI\AssignmentDE> docker cp 76893e740fa83821571483bb56da8c3259209b1af46384f38cdec63e8688ba94:/opt/hadoop/part-00000-efcbec96-5522-4a0f-852b-65404a81f32b-c000.csv .
Successfully copied 13.8kB to C:\Users\Administrator\Desktop\VDI\AssignmentDE\

```

Một số giá trị kết quả

1	6/11/2024,1,Mai Đức An,execute,3
2	6/12/2024,1,Mai Đức An,execute,10
3	6/15/2024,1,Mai Đức An,read,7
4	6/11/2024,1,Mai Đức An,read,11
5	6/10/2024,1,Mai Đức An,read,25
6	6/13/2024,1,Mai Đức An,read,14
7	6/14/2024,1,Mai Đức An,write,4
8	6/13/2024,1,Mai Đức An,write,10
9	6/10/2024,1,Mai Đức An,write,6
10	6/12/2024,2,Nguyễn Mai Anh,execute,1
11	6/13/2024,2,Nguyễn Mai Anh,read,3
12	6/15/2024,2,Nguyễn Mai Anh,write,2
13	6/11/2024,2,Nguyễn Mai Anh,write,1
14	6/12/2024,2,Nguyễn Mai Anh,write,19
15	6/12/2024,3,Ngô Ngọc Tuấn Anh,execute,9
16	6/10/2024,3,Ngô Ngọc Tuấn Anh,execute,4
17	6/13/2024,3,Ngô Ngọc Tuấn Anh,execute,9
18	6/13/2024,3,Ngô Ngọc Tuấn Anh,read,18
19	6/15/2024,3,Ngô Ngọc Tuấn Anh,read,24
20	6/13/2024,3,Ngô Ngọc Tuấn Anh,write,3
21	6/12/2024,3,Ngô Ngọc Tuấn Anh,write,2
22	6/10/2024,3,Ngô Ngọc Tuấn Anh,write,8
23	6/11/2024,4,Trần Trung Anh,execute,7

Link source code github:

<https://github.com/PJavis/AssignmentDE.git>