

Project: Building a Data Ingestion Pipeline with Apache Sqoop, Flume, and Kafka on a Local Machine

Part 1: Data Migration with Apache Sqoop

- Overview

This part of project demonstrates data migration from a local relational database (MySQL or SQLite) to Hadoop's HDFS using Apache Sqoop. The process involves setting up a local database, populating it with sample data, importing the data to HDFS, and configuring incremental imports to handle new data entries efficiently.

- Steps to Create the Local Database

1. Create a Database (labs):
2. Create a Table (authors) and Insert Sample Data:

- Create a target directory in HDFS to import table data into

Command: `$hdfs dfs -mkdir /mywarehouse`

- Import Data to HDFS with Sqoop

1. Full Data Import

- **Sqoop Command:**

```
$ sqoop import --connect jdbc:mysql://localhost/labs \  
--username student --password student \  
--table authors --fields-terminated-by '\t' \  
--target-dir /mywarehouse/authors_table
```

```
[student@192 ~]$ hdfs dfs -ls /mywarehouse/authors_table
Found 5 items
-rw-r--r-- 1 student supergroup 0 2024-08-08 08:30 /mywarehouse/authors_table/_SUCCESS
-rw-r--r-- 1 student supergroup 189746 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00000
-rw-r--r-- 1 student supergroup 190603 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00001
-rw-r--r-- 1 student supergroup 190637 2024-08-08 08:30 /mywarehouse/authors_table/part-m-00002
-rw-r--r-- 1 student supergroup 190593 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00003
[student@192 ~]$
```

2. Incremental Data Import

Note: I inserted new rows in 'authors' table for testing

- **Sqoop Command:**

```
$ sqoop import --connect jdbc:mysql://localhost/labs \
--username student --password student \
--table authors --fields-terminated-by '\t' \
--target-dir /mywarehouse/authors_table \
--incremental append \
--check-column id --last-value
```

- Validation and Testing

- **Verify Data in HDFS**

```
deleted /mywarehouse/authors_table
[student@192 ~]$ hdfs dfs -ls /mywarehouse/authors_table
Found 5 items
-rw-r--r-- 1 student supergroup 0 2024-08-08 08:30 /mywarehouse/authors_table/_SUCCESS
-rw-r--r-- 1 student supergroup 189746 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00000
-rw-r--r-- 1 student supergroup 190603 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00001
-rw-r--r-- 1 student supergroup 190637 2024-08-08 08:30 /mywarehouse/authors_table/part-m-00002
-rw-r--r-- 1 student supergroup 190593 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00003
[student@192 ~]$ hdfs dfs -ls /mywarehouse/authors_table
Found 9 items
-rw-r--r-- 1 student supergroup 0 2024-08-08 08:30 /mywarehouse/authors_table/_SUCCESS
-rw-r--r-- 1 student supergroup 189746 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00000
-rw-r--r-- 1 student supergroup 190603 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00001
-rw-r--r-- 1 student supergroup 190637 2024-08-08 08:30 /mywarehouse/authors_table/part-m-00002
-rw-r--r-- 1 student supergroup 190593 2024-08-08 08:29 /mywarehouse/authors_table/part-m-00003
-rw-r--r-- 1 student student 229 2024-08-08 08:38 /mywarehouse/authors_table/part-m-00004
-rw-r--r-- 1 student student 152 2024-08-08 08:39 /mywarehouse/authors_table/part-m-00005
-rw-r--r-- 1 student student 168 2024-08-08 08:39 /mywarehouse/authors_table/part-m-00006
-rw-r--r-- 1 student student 235 2024-08-08 08:39 /mywarehouse/authors_table/part-m-00007
[student@192 ~]$
```

- Future Enhancements

- **Automate Incremental Imports:** Use a cron job or another scheduler to automate the incremental import process.
- **Data Transformation:** Implement data transformation using Apache Hive or Apache Spark before storing the data in HDFS.

Part 2: Real-Time Data Ingestion with Apache Flume and Apache Kafka

- Overview

This project demonstrates the setup and configuration of a real-time data ingestion pipeline using Apache Flume and Apache Kafka. The objective is to capture log data from a local directory using Flume, send it to a Kafka topic, and store/manage the incoming log data using Kafka.

- Environment Preparation

1. Stop any running services (HBase, Kafka, Zookeeper):

```
sudo stop-hbase.sh  
sudo systemctl stop kafka  
sudo systemctl stop zookeeper
```

2. Start Zookeeper and Kafka:

```
sudo systemctl start zookeeper  
sudo systemctl start kafka
```

3. Configure Apache Flume

```
agent1.sources = streaming-txt-source  
agent1.sinks = kafka-sink logger-sink  
agent1.channels = memory-channel  
agent1.sources.streaming-txt-source.type = spooldir  
agent1.sources.streaming-txt-source.spoolDir = /home/student/Labs/C3U4/spool  
agent1.sinks.kafka-sink.type = org.apache.flume.sink.kafka.KafkaSink  
agent1.sinks.kafka-sink.topic = stream_text  
agent1.sinks.kafka-sink.brokerList = localhost:9092  
agent1.sinks.kafka-sink.batchSize = 5  
agent1.channels.memory-channel.type = memory  
agent1.channels.memory-channel.capacity = 10000  
agent1.channels.memory-channel.transactionCapacity = 100  
agent1.sinks.logger-sink.type = logger  
agent1.sources.streaming-txt-source.channels = memory-channel  
agent1.sinks.kafka-sink.channel = memory-channel  
agent1.sinks.logger-sink.channel = memory-channel
```

4. Configure Apache Kafka

- Create a new Kafka topic:
Command:
kafka-topics --create \
--bootstrap-server localhost:9092 \
--replication-factor 1 \
--partitions 1 \
--topic stream_text

5. Prepare the Spool Directory

```
cd /home/student/Labs/C3U4  
mkdir spool
```

- Running the Real-Time Data Ingestion Pipeline

1. Start the Flume Agent

```
flume-ng agent --conf $FLUME_HOME/conf \  
--conf-file /home/student/Labs/C3U4/spooldir.conf \  
--name agent1 -Dflume.root.logger=INFO,console
```

2. Generate Log Data

- Open another terminal and run a script to generate log data:

```
python ./spool_stream.py ./spool 5000 ~/Data/alice_in_wonderland.txt
```

3. Consume Data from Kafka

- Open a new terminal (consumer) and consume the data from the Kafka topic:

```
kafka-console-consumer \  
--bootstrap-server localhost:9092 \  
--topic stream_text \  
--from-beginning
```

- Validation and Testing

- **Check Kafka Topic:** Ensure that data is being consumed by the Kafka topic stream_text.
- **Monitor Flume Logs:** Check Flume logs to ensure that data is being captured and forwarded correctly.

The image shows three terminal windows on a dark background. The leftmost window, titled 'C3U4: python - Konsole', shows the command prompt where the user navigates to a directory and runs a Python script. The middle window, titled 'producer - Konsole', displays the output of the producer script, which includes timestamps, log levels, and the body of the data being sent. The rightmost window, titled 'consumer - Konsole', displays the output of the consumer script, which shows the received data as text from a story.

```
C3U4: python - Konsole
File Edit View Bookmarks Settings Help
[student@192 ~]$ cd /home/student/Labs/C3U4
[student@192 C3U4]$ ls
spool spooldir.conf spool_stream.py
[student@192 C3U4]$ python ./spool_stream.py ./
derland.txt
Press Ctrl+C
^C
```

```
producer - Konsole
File Edit View Bookmarks Settings Help
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: 4F 6E 20 77 68 69 63 68 20 53 65 76 65 6E 20 6C On which Seven l }
2024-08-09 05:51:03,181 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: 62 6C 61 60 65 20 6F 6E 20 6F 74 68 65 72 73 21 blame on others! }
2024-08-09 05:51:03,181 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: }
2024-08-09 05:51:03,181 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: E2 80 9C 5F 59 6F 75 E2 80 99 64 5F 20 62 65 74 ... You...d bet }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: 79 65 73 74 65 72 64 61 79 20 79 6F 75 20 64 65 yesterday you de }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: E2 80 9C 57 68 61 74 20 66 6F 72 3F E2 80 9D 20 ...What for?... }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: E2 80 9C 54 68 61 74 E2 80 99 73 20 6E 6F 65 ...That...s none }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: }
2024-08-09 05:51:03,182 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)) Event: { headers:{
} body: E2 80 9C 59 65 73 2C 20 69 74 20 5F 69 73 5F 20 ...Yes, it is_ }
^C
```

```
consumer - Konsole
File Edit View Bookmarks Settings Help
"yes, but some crumbs must have got in as well," the Hatter grumbled:
"you shouldn't have put it in with the bread-knife."

The March Hare took the watch and looked at it gloomily: then he dipped

"The Dormouse is asleep again," said the Hatter, and he poured a little
hot tea upon its
nose.

The Dormouse shook its head impatiently, and said, without opening its
eyes, "Of course, of course; just what I was going to remark myself."

The Hatter shook his head mournfully. "Not I!" he replied. "We
How I wonder what you're at!'

"I'm afraid I don't know one," said Alice, rather alarmed at the
hoarse, feeble voice: "I heard every word you fellows were saying."

"There's no such thing!" Alice was beginning very angrily, but the
Hatter and the March Hare went "Sh! sh!" and the Dormouse sulkily
remarked, "If you can't be civil, you'd better finish the story for
yourself."

everything's curious today. I think I may as well go in at once." And
CHAPTER VIII.

^C
```