

# CS4830 : Big Data Lab

## Lab 7 - Report

Divya K Raman EE15B085 and Sahana Ramnath EE15B109

April 2, 2019

The **python files** used were :

- lab7\_gettcp.py : Loads iris.csv from disk and sens each row to “GetTCP” processor in nifi
- lab7\_kafka.py : Accepts each row of iris data sent by “PublishKafka” processor in nifi, predicts its class using the saved trained model ‘finalized\_model.sav’ and sends the features and predicted data to “ConsumerKafka” processor in nifi.
- lab7\_puttcp.py : Accepts (feature,prediction) rows from PutTCP and print them

These can be found in the submitted folder.

**Commands** to be run :

- cd kafka\_2.1.1-2.10
- bin/zookeeper-server-start.sh config/zookeeper.properties
- bin/kafka-server-start.sh config/server.properties
- bin/kafka-topics.sh --create --zookeeper localhost:9092 --replication-factor 1 --partitions 2 --topic virginicasetosa
- bin/kafka-topics.sh --create --zookeeper localhost:9092 --replication-factor 1 --partitions 2 --topic versicolor
- python lab7\_gettcp.py, python lab7\_kafka.py, python lab7\_puttcp.py (in 3 terminals in parallel)

**Nifi Architecture :**

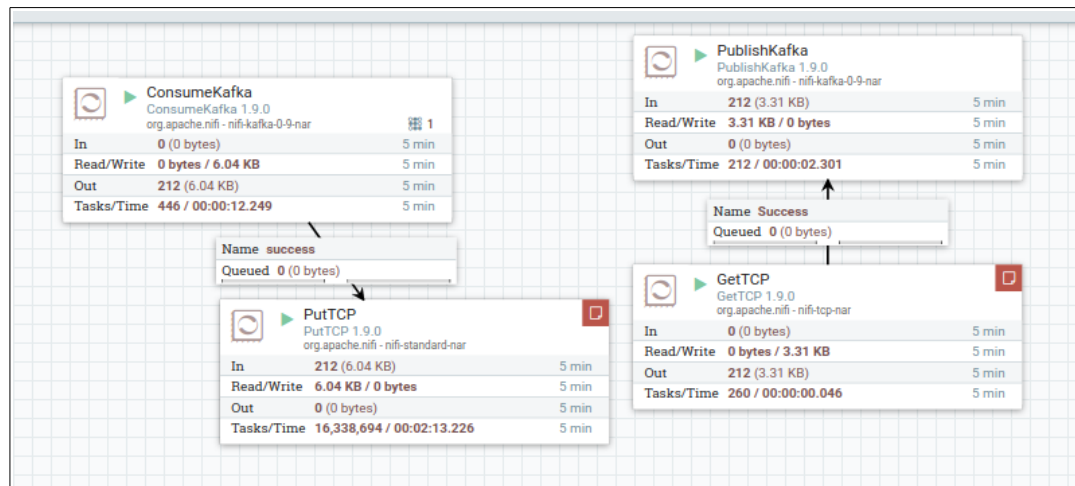
Node 2 : GetTCP -----> PublisherKafka

Node 1 : ConsumerKafka -----> PutTCP

Note : ValidateRecords threw errors and didn't allow flow of data and hence was omitted in the architecture used.

Configuration and architecture are shown in the figures below.

### NIFI ARCHITECTURE



The data flow can be seen in the above figure.

GetTCP Configuration

Processor Details

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value
Endpoint List	127.0.0.1:9081
Connection Attempt Count	10
Reconnect interval	5 sec
Receive Buffer Size	16MB
End of message delimiter byte	10

OK

PublisherKafka Configuration

Processor Details

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property	Value
Kafka Brokers	localhost:9092
Security Protocol	PLAINTEXT
Kerberos Service Name	No value set
SSL Context Service	No value set
Topic Name	virginicasetosa
Delivery Guarantee	Best Effort
Kafka Key	No value set
Key Attribute Encoding	UTF-8 Encoded
Message Demarcator	No value set
Max Request Size	1 MB
Acknowledgment Wait Time	5 secs
Max Metadata Wait Time	5 sec
Partitioner class	DefaultPartitioner
Compression Type	none

OK

## ConsumerKafka Configuration

### Processor Details

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property		Value	
Kafka Brokers	?	localhost:9092	
Security Protocol	?	PLAINTEXT	
Kerberos Service Name	?	No value set	
SSL Context Service	?	No value set	
Topic Name(s)	?	versicolor	
Group ID	?	0	
Offset Reset	?	latest	
Key Attribute Encoding	?	UTF-8 Encoded	
Message Demarcator	?	No value set	
Max Poll Records	?	10000	
Max Uncommitted Time	?	1 secs	

OK

## PutTCP Configuration

### Processor Details

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

Property		Value	
Hostname	?	localhost	
Port	?	5002	
Max Size of Socket Send Buffer	?	1 MB	
Idle Connection Expiration	?	5 seconds	
Connection Per FlowFile	?	false	
Outgoing Message Delimiter	?	No value set	
Timeout	?	10 seconds	
SSL Context Service	?	No value set	
Character Set	?	UTF-8	

OK

## SCREENSHOTS OF OUTPUTS

```
sahana@sahana-Inspiron-5559:/media/sahana/48BC3293BC327C0E/Sahan/Sem8/BigDataLab/EE15B085_EE15B109_Lab7$ python lab7_gettcp.py
('Connection address:', ('127.0.0.1', 55580))
5.1 3.5 1.4 0.2 Iris-setosa
0 4.9 3.0 1.4 0.2 Iris-setosa
1 4.7 3.2 1.3 0.2 Iris-setosa
2 4.6 3.1 1.5 0.2 Iris-setosa
3 5.0 3.6 1.4 0.2 Iris-setosa
4 5.4 3.9 1.7 0.4 Iris-setosa
Message sent to GetTCP : 4.9,3.0,1.4,0.2
Message sent to GetTCP : 4.7,3.2,1.3,0.2
Message sent to GetTCP : 4.6,3.1,1.5,0.2
Message sent to GetTCP : 5.0,3.6,1.4,0.2
Message sent to GetTCP : 5.4,3.9,1.7,0.4
Message sent to GetTCP : 4.6,3.4,1.4,0.3
Message sent to GetTCP : 5.0,3.4,1.5,0.2
Message sent to GetTCP : 4.4,2.9,1.4,0.2
Message sent to GetTCP : 4.9,3.1,1.5,0.1
Message sent to GetTCP : 5.4,3.7,1.5,0.2
Message sent to GetTCP : 4.8,3.4,1.6,0.2
```

```
sahana@sahana-Inspiron-5559:/media/sahana/48BC3293BC327C0E/Sahan/Sem8/BigDataLab/EE15B085_EE15B109_Lab7$ python lab7_kafka.py
Before receiving..
Received message from PublishKafka is : 4.9,3.0,1.4,0.2
Message sent to ConsumerKafka : Iris-setosa
Received message from PublishKafka is : 4.7,3.2,1.3,0.2
Message sent to ConsumerKafka : Iris-setosa
Received message from PublishKafka is : 4.6,3.1,1.5,0.2
Message sent to ConsumerKafka : Iris-setosa
Received message from PublishKafka is : 5.0,3.6,1.4,0.2
Message sent to ConsumerKafka : Iris-setosa
Received message from PublishKafka is : 5.4,3.9,1.7,0.4
Message sent to ConsumerKafka : Iris-setosa
Received message from PublishKafka is : 4.6,3.4,1.4,0.3
Message sent to ConsumerKafka : Iris-setosa
Received message from PublishKafka is : 5.0,3.4,1.5,0.2
Message sent to ConsumerKafka : Iris-setosa
```

```
sahana@sahana-Inspiron-5559:/media/sahana/48BC3293BC327C0E/Sahan/Sem8/BigDataLab/EE15B085_EE15B109_Lab7$ python lab7_puttcp.py
('Connection address:', ('127.0.0.1', 34646))
Message received from PutTCP : 4.7,3.2,1.3,0.2 Iris-setosa
Message received from PutTCP : 4.9,3.0,1.4,0.2 Iris-setosa
Message received from PutTCP : 4.6,3.1,1.5,0.2 Iris-setosa
Message received from PutTCP : 5.0,3.6,1.4,0.2 Iris-setosa
Message received from PutTCP : 5.4,3.9,1.7,0.4 Iris-setosa
Message received from PutTCP : 4.6,3.4,1.4,0.3 Iris-setosa
Message received from PutTCP : 5.0,3.4,1.5,0.2 Iris-setosa
Message received from PutTCP : 4.4,2.9,1.4,0.2 Iris-setosa
Message received from PutTCP : 4.9,3.1,1.5,0.1 Iris-setosa
Message received from PutTCP : 5.4,3.7,1.5,0.2 Iris-setosa
Message received from PutTCP : 4.8,3.4,1.6,0.2 Iris-setosa
Message received from PutTCP : 4.8,3.0,1.4,0.1 Iris-setosa
Message received from PutTCP : 4.3,3.0,1.1,0.1 Iris-setosa
Message received from PutTCP : 5.8,4.0,1.2,0.2 Iris-setosa
Message received from PutTCP : 5.7,4.4,1.5,0.4 Iris-setosa
Message received from PutTCP : 5.4,3.9,1.3,0.4 Iris-setosa
Message received from PutTCP : 5.1,3.5,1.4,0.3 Iris-setosa
Message received from PutTCP : 5.7,3.8,1.7,0.3 Iris-setosa
Message received from PutTCP : 5.1,3.8,1.5,0.3 Iris-setosa
Message received from PutTCP : 5.4,3.4,1.7,0.2 Iris-setosa
Message received from PutTCP : 5.1,3.7,1.5,0.4 Iris-setosa
Message received from PutTCP : 4.6,3.6,1.0,0.2 Iris-setosa
Message received from PutTCP : 5.1,3.3,1.7,0.5 Iris-setosa
```

The above three screenshots are by lab7\_gettcp.py, lab7\_kafka.py and lab7\_puttcp.py respectively.

## **Functions used in Python :**

For kafka, KafkaProducer and KafkaConsumer were used.

Functions of module socket were used to send and receive data from nifi through TCP.

## **OTHER NIFI PROCESSORS**

1. GetHDFSFileInfo : Retrieves a listing of files and directories from HDFS. This processor creates a FlowFile(s) that represents the HDFS file/dir with relevant information. Main purpose of this processor to provide functionality similar to HDFS Client, i.e. count, ls, test, etc. This processor is stateless, supports incoming connections and provides information on a dir level.
2. HashAttribute : Hashes together the key/value pairs of several flowfile attributes and adds the hash as a new attribute. Optional properties are to be added such that the name of the property is the name of a flowfile attribute to consider and the value of the property is a regular expression that, if matched by the attribute value, will cause that attribute to be used as part of the hash. If the regular expression contains a capturing group, only the value of the capturing group will be used. For a processor which accepts various attributes and generates a cryptographic hash of each, see "CryptographicHashAttribute".
3. ListSFTP : Performs a listing of the files residing on an SFTP server. For each file that is found on the remote server, a new FlowFile will be created with the filename attribute set to the name of the file on the remote server. This can then be used in conjunction with FetchSFTP in order to fetch those files.

([This](#) link was referred to for the above).