**LABORATORY PROGRAM – 8**

**Scala Program to print numbers from 1 to 100**

from pyspark import SparkContext

from pyspark.streaming import StreamingContext

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer

from nltk.tokenize import word\_tokenize

import re

# Initialize SparkContext and StreamingContext (batch interval = 5 sec)

sc = SparkContext("local[2]", "TextCleanStreamingApp")

ssc = StreamingContext(sc, 5)

# Set up stop words and lemmatizer

stop\_words = set(stopwords.words('english'))

lemmatizer = WordNetLemmatizer()

# Connect to the socket stream

lines = ssc.socketTextStream("localhost", 9999)

# Text cleaning function

def clean\_text(line):

# Remove non-alphabetic characters and lower the case

line = re.sub(r'[^a-zA-Z\s]', '', line)

line = line.lower()

# Tokenize

tokens = word\_tokenize(line)

# Remove stop words and lemmatize

cleaned\_tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop\_words]

return ' '.join(cleaned\_tokens)

# Apply cleaning to each line

cleaned\_lines = lines.map(clean\_text)

# Print the cleaned lines

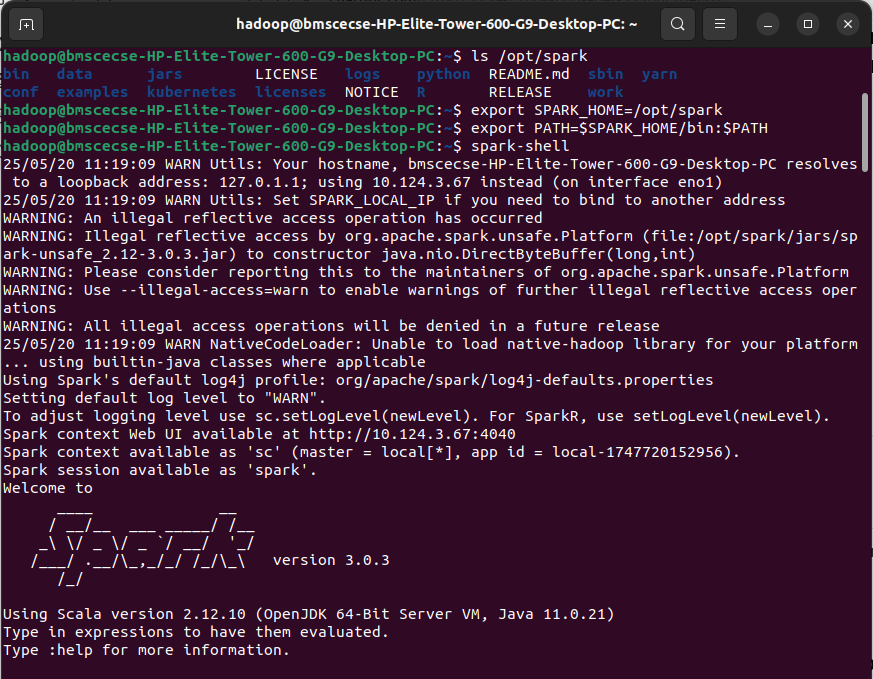
cleaned\_lines.pprint()

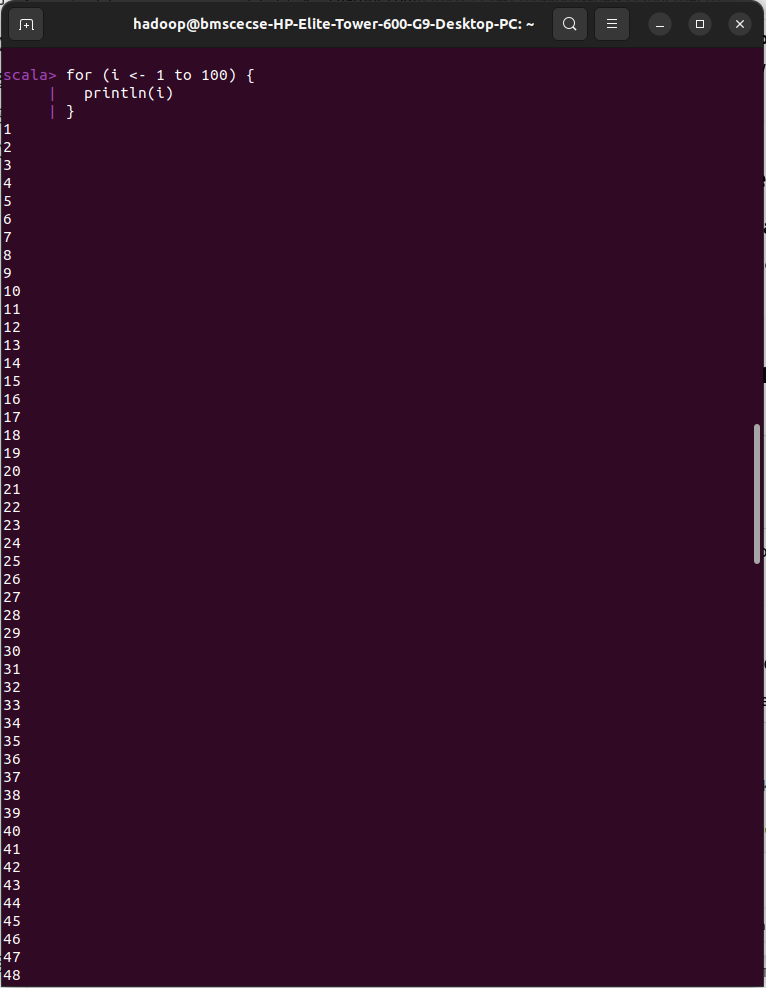
# Start streaming

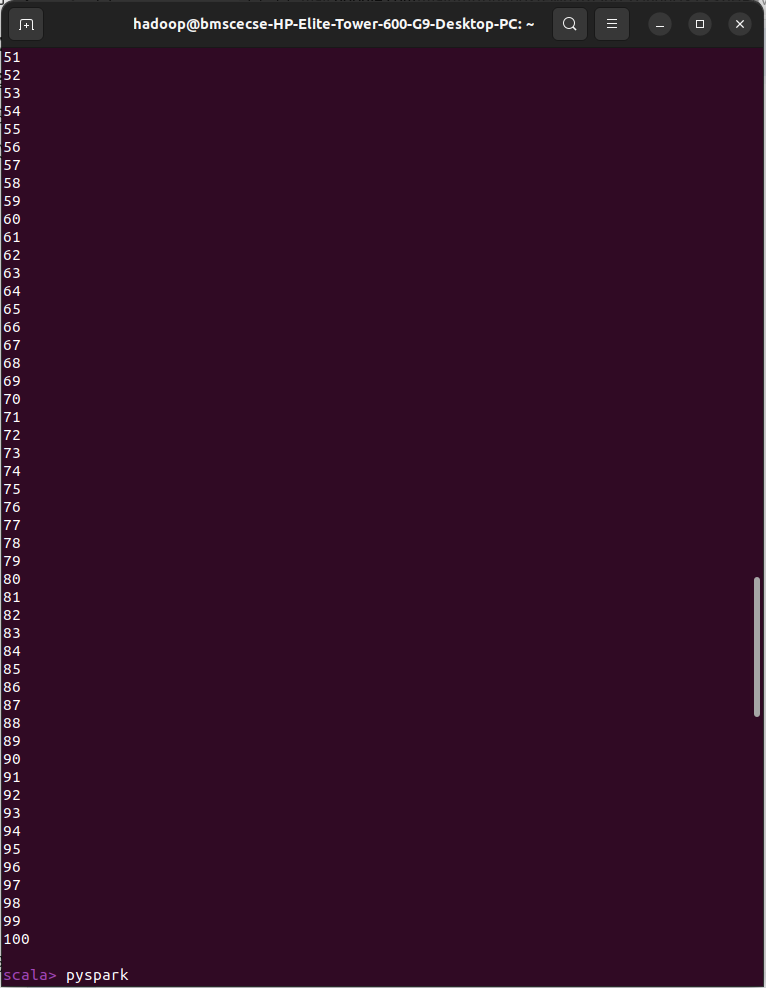
ssc.start()

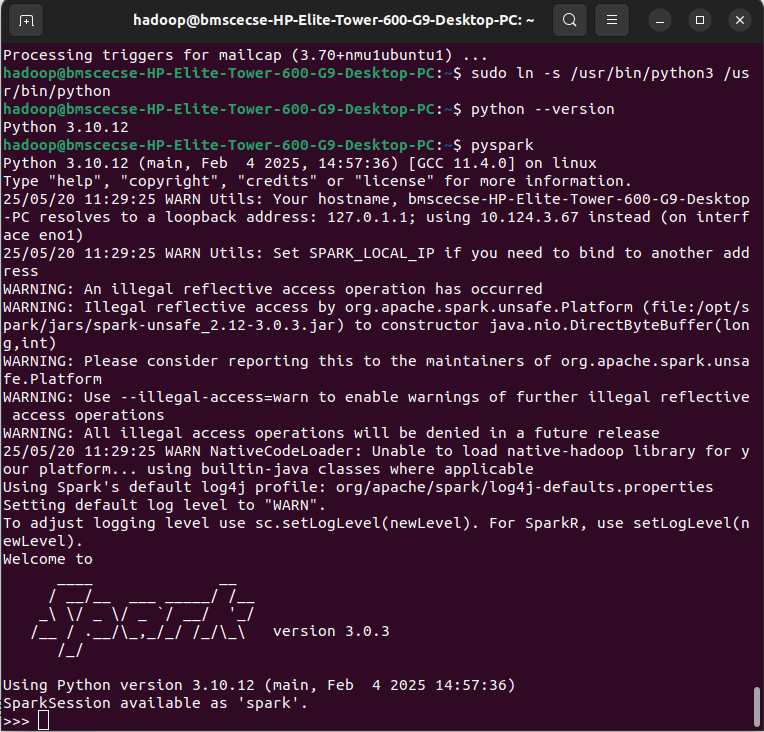
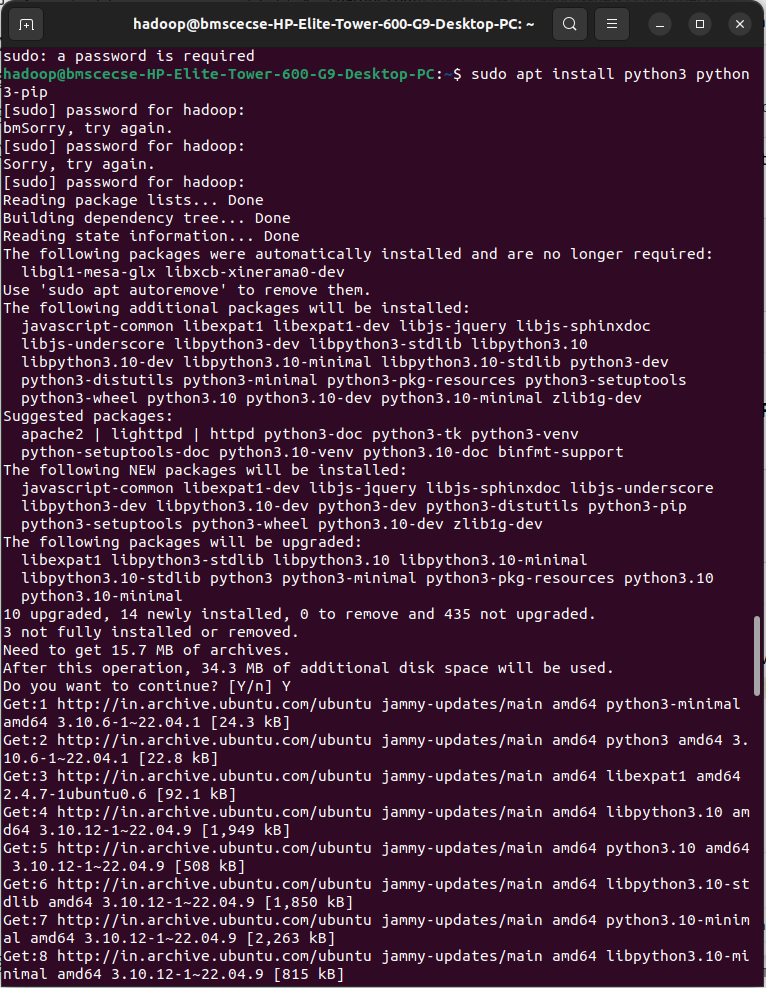
ssc.awaitTermination()

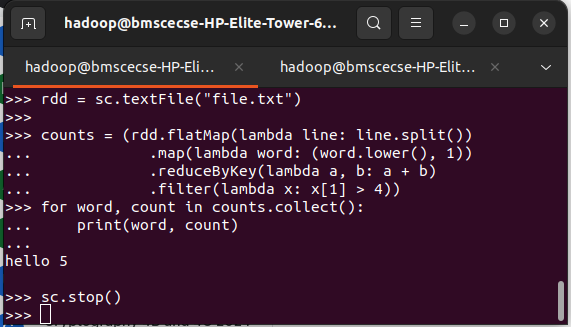
**OBSERVATION**

****

****

****

****

****