**Part 1:**

**Approach:**

1. Created a Twitter developer account and took elevated access for the account.
2. Created a new application in the account (CovidandOlympics).
3. After app creation obtained and stored all the keys.
4. Created a configuration file for the flume agent by specifying the consumer key, consumer secret, access token, access token secret, keywords and HDFS path. (Will update the contents of the .conf file in a PDF)
5. Created a new directory, to store the keyword-based data downloaded from twitter using flume.
6. Created a .conf file and moved it to the local of csx from the system.
7. Using flume commands collected the data based on both the keywords Covid and Olympics.
8. Total size of the data collected is 178MB.
9. Killed the processed used while running nohup commands after the collection of enough data.

**Instructions to run code and screenshots:**

1,1) For creating directory in Hadoop use following command:

- hdfs dfs –mkdir /user/ssamine/covid\_olympics\_dt (for covid and Olympics keyword data)

Text

Description automatically generated

Figure 1: Shows the creation of a directory

1,2) For collecting the data from twitter using Flume use the following command:

- nohup $FLUME\_HOME/bin/flume/-ng agent –n TwitterAgent -f/home/ssamine/covid\_olympics.conf –conf /usr/local/flume/conf &

Graphical user interface, text

Description automatically generated

Figure 2: Shows the command for data downloading from twitter using flume

1,3) To check if the data is downloading and the downloaded data size, use the following command:

- hdfs dfs –du –s –h /user/ssamine/covid\_olympics\_dt

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3: Shows the command to view the downloading data and size

1,4) To stop the nohup processes after collecting enough data use the following command:

- Kill -9 processid

(Used the command to kill all the processes running for my id: ‘pkill -KILL -u ssamine -f flume’, As the data download didn’t stop after killing the process id)

A screenshot of a computer

Description automatically generated with medium confidence

Figure 4: Shows the command to stop download by killing process ID

1,5) To check on the data downloaded:

- hdfs dfs -ls /user/ssamine/covid\_olympics\_dt/2022/03/06/23 ( year/month/date/time)

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Figure 5: Shows the command to view the downloaded data

1,6) To view the content of a file in HDFS that stores the tweets:

- hdfs dfs -cat /user/ssamine/covid\_olympics\_dt/2022/03/06/23/FlumeData.1646630307663 (a record id that was downloaded)

A screenshot of a computer

Description automatically generated with medium confidence

Figure 6: Shows the command to view the contents of a file in HDFS which stores tweets