Task 1

Part A

the command options poped up when I entered “flume-ng help”

```
eric@eric-VirtualBox:/usr/local/flume/bin$ flume-ng help
Usage: /usr/local/flume/bin/flume-ng <command> [options]...

commands:
  help                      display this help text
  agent                     run a Flume agent
  avro-client               run an avro Flume client
  version                   show Flume version info

global options:
  --conf,-c <conf>          use configs in <conf> directory
  --classpath,-C <cp>       append to the classpath
  --dryrun,-d               do not actually start Flume, just print the comman
d
  --plugins-path <dirs>     colon-separated list of plugins.d directories. See
 the
                            plugins.d section in the user guide for more detai
ls.
                            Default: $FLUME_HOME/plugins.d
  -Dproperty=value          sets a Java system property value
  -Xproperty=value          sets a Java -X option

agent options:
  --name,-n <name>          the name of this agent (required)
  --conf-file,-f <file>     specify a config file (required if -z missing)
  --zkConnString,-z <str>   specify the ZooKeeper connection to use (required
if -f missing)
  --zkBasePath,-p <path>    specify the base path in ZooKeeper for agent confi
gs
  --no-reload-conf          do not reload config file if changed
  --help,-h                 display help text

avro-client options:
  --rpcProps,-P <file>   RPC client properties file with server connection par
ams
  --host,-H <host>       hostname to which events will be sent
  --port,-p <port>       port of the avro source
  --dirname <dir>        directory to stream to avro source
  --filename,-F <file>   text file to stream to avro source (default: std inpu
t)
  --headerFile,-R <file> File containing event headers as key/value pairs on e
ach new line
  --help,-h              display help text

  Either --rpcProps or both --host and --port must be specified.

Note that if <conf> directory is specified, then it is always included first
in the classpath.
```

```
eric@eric-VirtualBox:/usr/local/flume/bin$ flume-ng version
Flume 1.9.0
Source code repository: https://git-wip-us.apache.org/repos/asf/flume.git
Revision: d4fcab4f501d41597bc616921329a4339f73585e
Compiled by fszabo on Mon Dec 17 20:45:25 CET 2018
From source with checksum 35db629a3bda49d23e9b3690c80737f9
```

Part B

The picture displays the contents displayed in the HDFS system, since I copied the files one by one and turned down the flume between, so the data are storied into two files.

```
eric@eric-VirtualBox:~$ hdfs dfs -ls /user/flume/
Found 2 items
-rw-r--r--   1 eric supergroup        404 2020-12-09 18:42 /user/flume/FlumeData
.1607557324328
-rw-r--r--   1 eric supergroup        274 2020-12-09 18:42 /user/flume/FlumeData
.1607557348690
eric@eric-VirtualBox:~$ hdfs dfs -cat /user/flume/FlumeData.1607557324328
Flume is a distributed, reliable, and available service for efficiently collecti
ng, aggregating, and moving large amounts of log data. It has a simple and flexi
ble architecture based on streaming data flows. It is robust and fault tolerant
with tunable reliability mechanisms and many failover and recovery mechanisms. I
t uses a simple extensible data model that allows for online analytic applicatio
n.

eric@eric-VirtualBox:~$ hdfs dfs -cat /user/flume/FlumeData.1607557348690
The Apache Hive data warehouse software facilitates reading, writing, and managi
ng large datasets residing in distributed storage using SQL. Structure can be pr
ojected onto data already in storage. A command line tool and JDBC driver are pr
ovided to connect users to Hive.
```

Task 2

Part A

```
eric@eric-VirtualBox:/usr/local/hive/conf$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.6.2.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4
j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-com
mon-2.3.7.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
```

Part B

We can see the average score by each subject has already displayed as the final result.

```
hive> select subject, avg(score) from scores GROUP BY subject;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the futu
re versions. Consider using a different execution engine (i.e. spark, tez) or us
ing Hive 1.X releases.
Query ID = eric_20201210222603_2ebfbde0-2b66-41eb-9127-02f88843396a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1607634711965_0001, Tracking URL = http://eric-VirtualBox:808
8/proxy/application_1607634711965_0001/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1607634711965_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-12-10 22:26:09,621 Stage-1 map = 0%,  reduce = 0%
2020-12-10 22:26:13,717 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.8 sec
2020-12-10 22:26:17,810 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.62
sec
MapReduce Total cumulative CPU time: 1 seconds 620 msec
Ended Job = job_1607634711965_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.62 sec   HDFS Read: 9406 HD
FS Write: 178 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 620 msec
OK
ece354  82.4375
ece571  78.6875
ece595  80.28571428571429
Time taken: 15.321 seconds, Fetched: 3 row(s)
```

Task 3

Part A

The Spark banner has already displayed in the terminal

```
eric@eric-VirtualBox:~$ pyspark
Python 3.8.5 (default, Jul 28 2020, 12:59:40)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/spark/jars/slf4j-log4j12-1.7.30.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4
j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
20/12/05 22:44:49 WARN util.Utils: Your hostname, eric-VirtualBox resolves to a
loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
20/12/05 22:44:49 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to ano
ther address
20/12/05 22:44:49 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.0.1
      /_/

Using Python version 3.8.5 (default, Jul 28 2020 12:59:40)
SparkSession available as 'spark'.
```

Part B

The result below shows the top 10 most frequent words and their frequencies in the “words.txt” file

```
20/12/11 01:55:11 INFO scheduler.DAGScheduler: Job 0 finished: takeOrdered at /home/eric/Desktop/Zihan_Wei_task3b.py:33, to
ok 1.095968 s
the     :       41
of      :       30
and     :       17
is      :       12
to      :       10
a       :       9
science         :       8
in      :       6
new     :       6
matter  :       6
```

Part C

The contents below show the “City Name” “Average Number” and “Standard Deviation” in order, since the result is pretty lengthy, so I screenshotted the result as much as I can.

```
North Las Vegas:        263.3, 154.914
Phoenix:        254.71, 144.113
Omaha:  274.8, 145.648
Anchorage:      242.33, 138.434
Anaheim:        267.54, 143.188
Greensboro:     282.5, 146.095
Dallas: 270.74, 151.753
Oakland:        276.15, 135.192
Laredo: 249.54, 140.046
Scottsdale:     274.61, 148.542
San Antonio:    272.73, 134.613
Bakersfield:    253.53, 162.943
Raleigh:        298.71, 143.802
Chula Vista:    216.2, 155.807
Philadelphia:   262.54, 131.246
Louisville:     223.46, 134.604
Los Angeles:    247.55, 154.332
Chandler:       239.07, 139.704
Sacramento:     260.18, 162.923
Indianapolis:   256.46, 142.739
Cleveland:      300.31, 144.406
San Diego:      231.5, 144.176
San Francisco:  249.5, 139.016
Nashville:      237.45, 152.354
Oklahoma City:  270.13, 146.405
Chesapeake:     270.95, 127.541
Detroit:        267.01, 163.615
Portland:       215.09, 149.439
Aurora: 281.24, 144.119
Boise:  249.36, 134.645
Baton Rouge:    249.36, 121.142
St. Louis:      254.8, 138.425
Birmingham:     238.73, 157.898
Plano:  236.75, 137.37
Irvine: 241.98, 134.095
Columbus:       274.15, 147.544
Memphis:        238.5, 147.484
Austin: 256.33, 133.849
Madison:        253.79, 158.631
Washington:     247.57, 125.398
El Paso:        242.58, 150.927
Milwaukee:      241.63, 149.643
Pittsburgh:     243.53, 166.498
Garland:        227.86, 141.329
Gilbert:        254.82, 156.673
Chicago:        273.96, 145.511
Lubbock:        283.6, 134.746
Tampa:  211.1, 133.116
Glendale:       276.56, 144.981
Newark: 306.53, 143.537
```