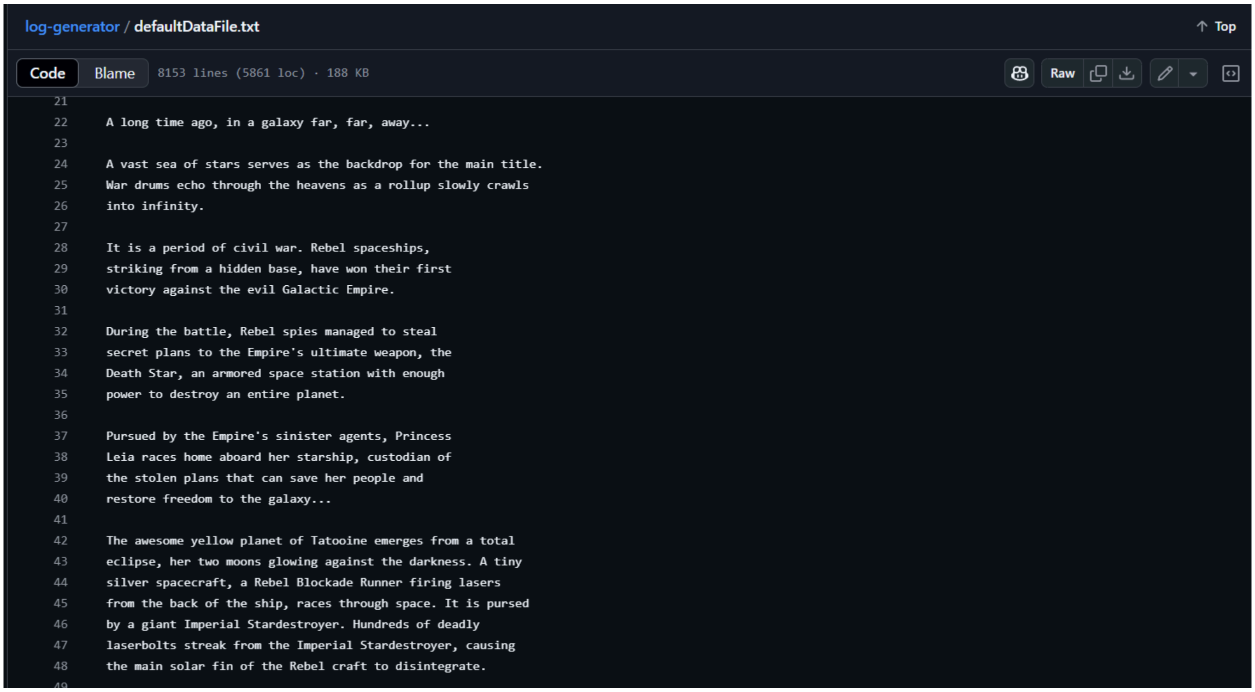
NTI Graduation Project

At the beginning, I generated logs using a Python script and a pre-existing data file.

Data Content:



```

A long time ago, in a galaxy far, far, away...

A vast sea of stars serves as the backdrop for the main title.
War drums echo through the heavens as a rollup slowly crawls
into infinity.

It is a period of civil war. Rebel spaceships,
striking from a hidden base, have won their first
victory against the evil Galactic Empire.

During the battle, Rebel spies managed to steal
secret plans to the Empire's ultimate weapon, the
Death Star, an armored space station with enough
power to destroy an entire planet.

Pursued by the Empire's sinister agents, Princess
Leia races home aboard her starship, custodian of
the stolen plans that can save her people and
restore freedom to the galaxy...

The awesome yellow planet of Tatooine emerges from a total
eclipse, her two moons glowing against the darkness. A tiny
silver spacecraft, a Rebel Blockade Runner firing lasers
from the back of the ship, races through space. It is pursed
by a giant Imperial Stardestroyer. Hundreds of deadly
laserbolts streak from the Imperial Stardestroyer, causing
the main solar fin of the Rebel craft to disintegrate.
```

Script:

log-generator

Simple python script to generate random time series data into a log file (GMT time) useful for generating sample time-series based data to feed other downstream systems.

Usage

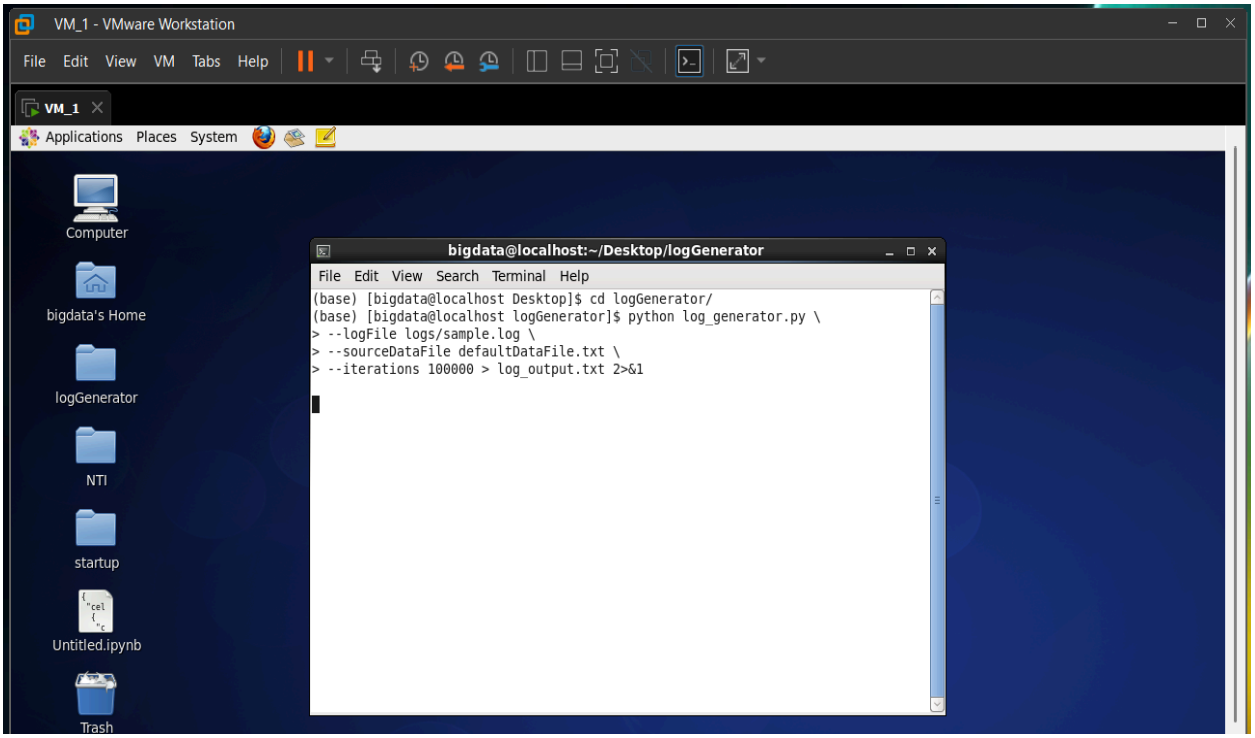
```
logGenerator.py --logFile <targetFile>
    [--minSleepMs <int>] [--maxSleepMs <int>]
    [--sourceDataFile <fileWithTextData>] [--iterations <long>]
    [--minLines <int>] [--maxLines <int>]
    [--logPattern <pattern>] [--datePattern <pattern>]
```

Defaults

```
iterations = -1 # infinite
minSleep = 0.1
maxSleep = 1
minLines = 1
maxLines = 1
logFile = 'logGenerator.log'
sourceDataFile = 'defaultDataFile.txt'
logPattern = '%(asctime)s,%(msecs)d %(process)d %(filename)s %(lineno)d %(name)s %(levelname)s %(messag
datePattern = "%Y-%m-%d %H:%M:%S"
```

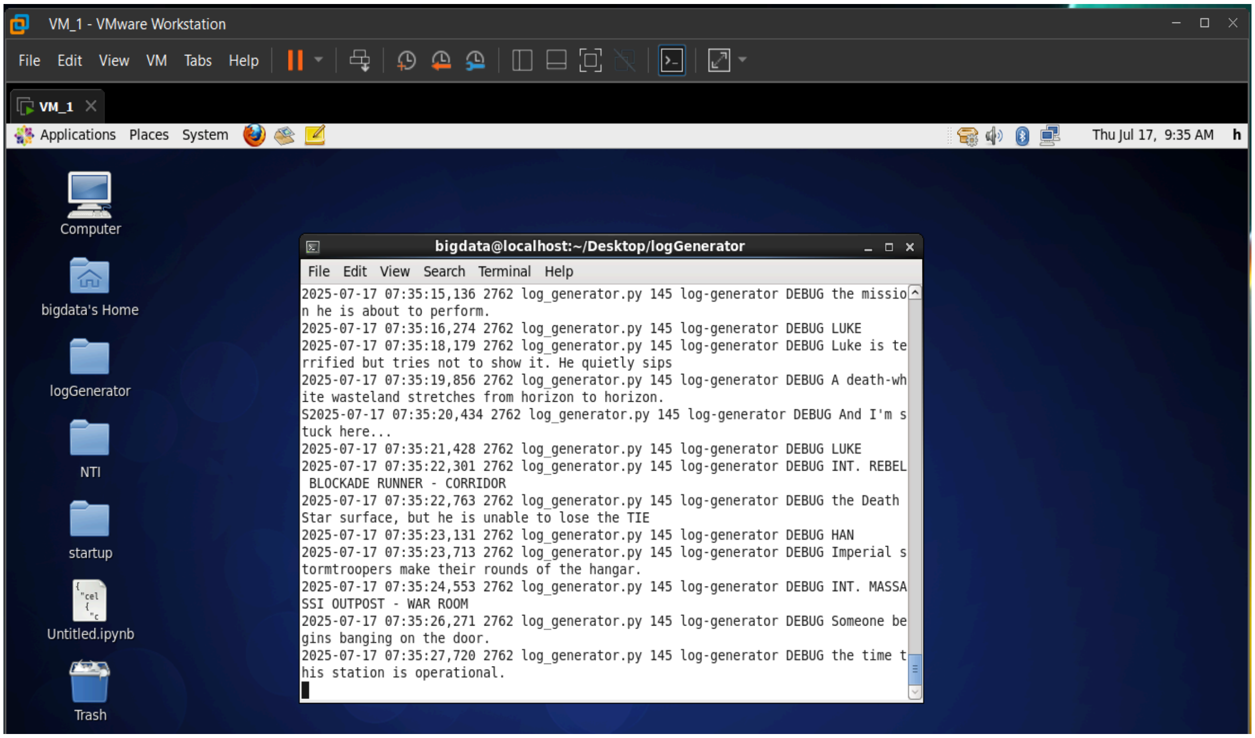
At the beginning, I ran the script using the following command:

Note: the log generator folder on the Desktop.



```
(base) [bigdata@localhost Desktop]$ cd logGenerator/
(base) [bigdata@localhost logGenerator]$ python log_generator.py \
> --logFile logs/sample.log \
> --sourceDataFile defaultDataFile.txt \
> --iterations 100000 > log_output.txt 2>&1
```

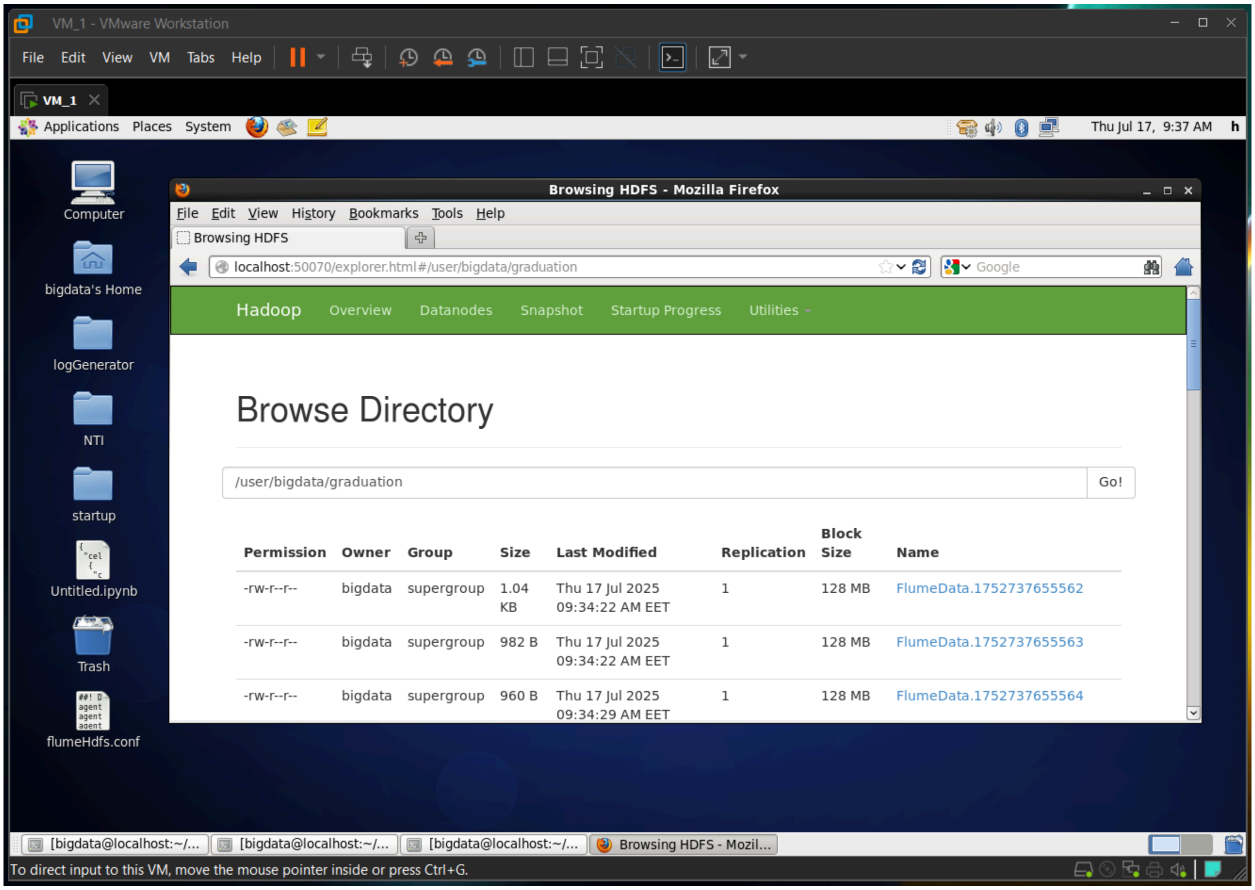
After that, I streamed the log data live using: [tail -f logs/sample.log]



```
2025-07-17 07:35:15,136 2762 log_generator.py 145 log-generator DEBUG the missio
n he is about to perform.
2025-07-17 07:35:16,274 2762 log_generator.py 145 log-generator DEBUG LUKE
2025-07-17 07:35:18,179 2762 log_generator.py 145 log-generator DEBUG Luke is te
rrified but tries not to show it. He quietly sips
2025-07-17 07:35:19,856 2762 log_generator.py 145 log-generator DEBUG A death-wh
ite wasteland stretches from horizon to horizon.
S2025-07-17 07:35:20,434 2762 log_generator.py 145 log-generator DEBUG And I'm s
tuck here...
2025-07-17 07:35:21,428 2762 log_generator.py 145 log-generator DEBUG LUKE
2025-07-17 07:35:22,301 2762 log_generator.py 145 log-generator DEBUG INT. REBEL
 BLOCKADE RUNNER - CORRIDOR
2025-07-17 07:35:22,763 2762 log_generator.py 145 log-generator DEBUG the Death
Star surface, but he is unable to lose the TIE
2025-07-17 07:35:23,131 2762 log_generator.py 145 log-generator DEBUG HAN
2025-07-17 07:35:23,713 2762 log_generator.py 145 log-generator DEBUG Imperial s
tormtroopers make their rounds of the hangar.
2025-07-17 07:35:24,553 2762 log_generator.py 145 log-generator DEBUG INT. MASSA
SSI OUTPOST - WAR ROOM
2025-07-17 07:35:26,271 2762 log_generator.py 145 log-generator DEBUG Someone be
gins banging on the door.
2025-07-17 07:35:27,720 2762 log_generator.py 145 log-generator DEBUG the time t
his station is operational.
```

After I started HDFS and created a directory called 'graduation', the logs were saved there using Flume.

```
hdfs dfs -mkdir -p /user/bigdata/graduation
```

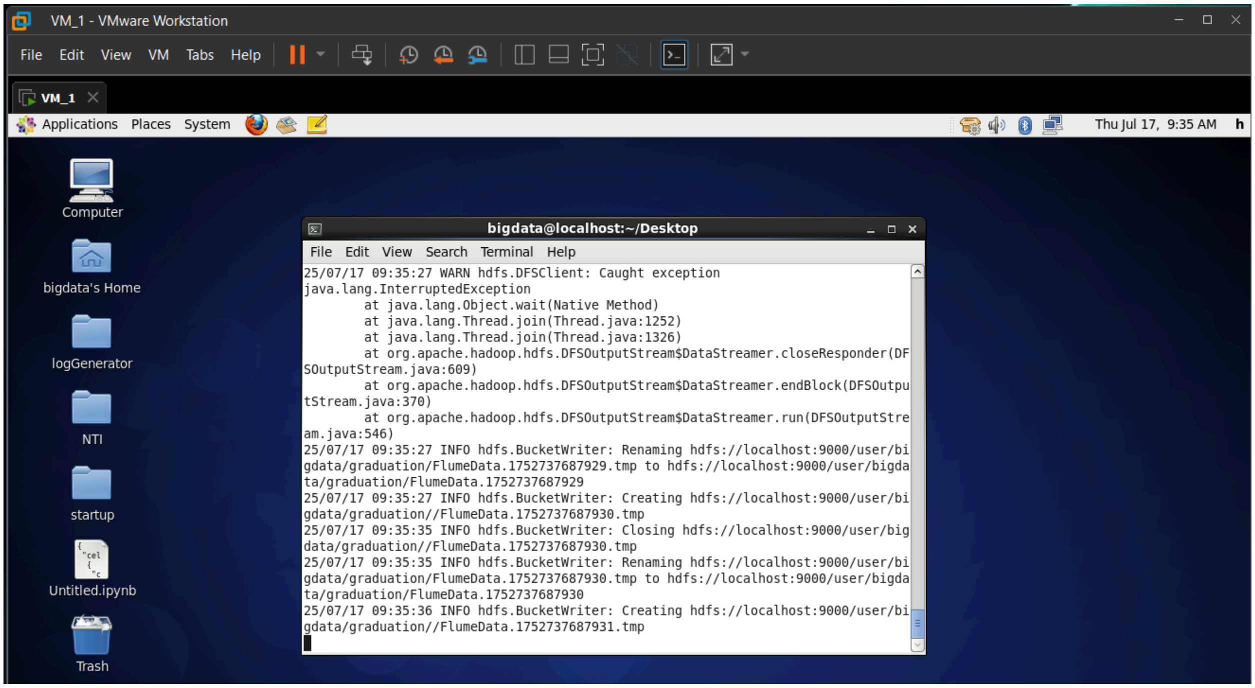


/user/bigdata/graduation

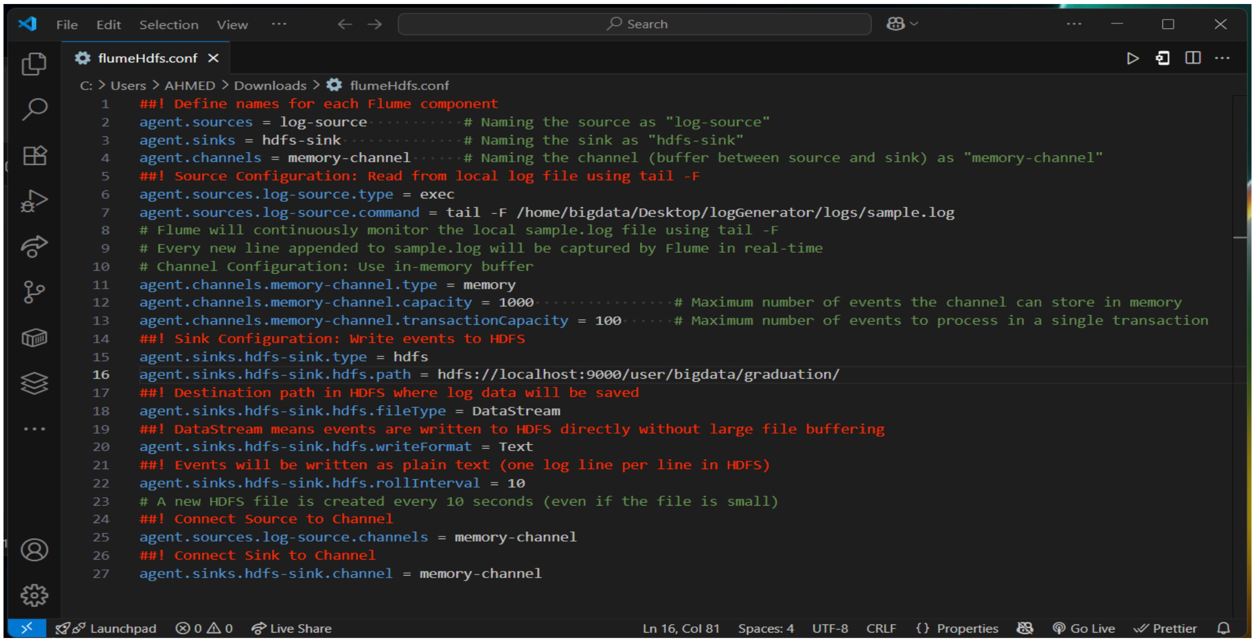
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	bigdata	supergroup	1.04 KB	Thu 17 Jul 2025 09:34:22 AM EET	1	128 MB	FlumeData.1752737655562
-rw-r--r--	bigdata	supergroup	982 B	Thu 17 Jul 2025 09:34:22 AM EET	1	128 MB	FlumeData.1752737655563
-rw-r--r--	bigdata	supergroup	960 B	Thu 17 Jul 2025 09:34:29 AM EET	1	128 MB	FlumeData.1752737655564

Then, I ran Flume with the following command after creating the configuration file:

```
flume-ng agent \
--conf conf \
--conf-file /home/bigdata/Desktop/flumeHdfs.conf \
--name agent \
-Dflume.root.logger=INFO,console
```

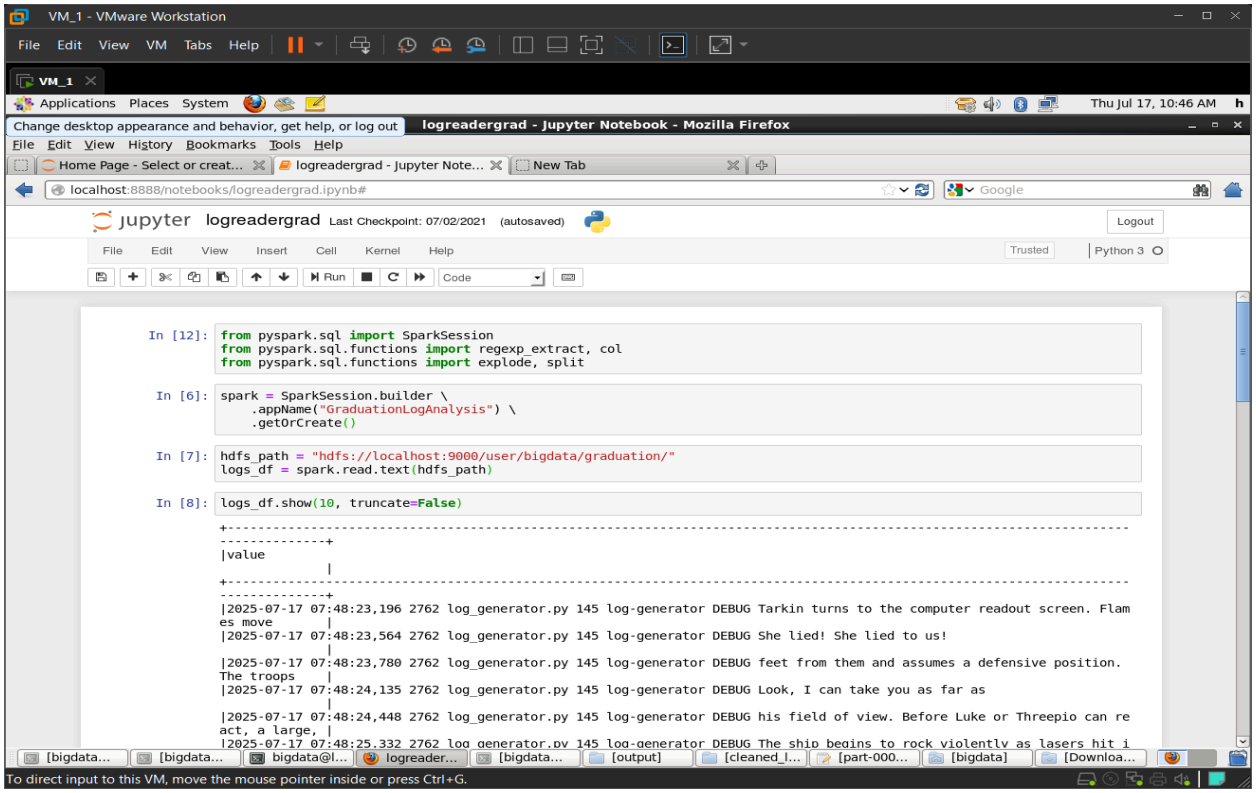


```
25/07/17 09:35:27 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:609)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:370)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:546)
25/07/17 09:35:27 INFO hdfs.BucketWriter: Renaming hdfs://localhost:9000/user/bi
gdata/graduation/FlumeData.1752737687929.tmp to hdfs://localhost:9000/user/bigda
ta/graduation/FlumeData.1752737687929
25/07/17 09:35:27 INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/bi
gdata/graduation//FlumeData.1752737687930.tmp
25/07/17 09:35:35 INFO hdfs.BucketWriter: Closing hdfs://localhost:9000/user/big
data/graduation//FlumeData.1752737687930.tmp
25/07/17 09:35:35 INFO hdfs.BucketWriter: Renaming hdfs://localhost:9000/user/bi
gdata/graduation/FlumeData.1752737687930.tmp to hdfs://localhost:9000/user/bigda
ta/graduation/FlumeData.1752737687930
25/07/17 09:35:36 INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/bi
gdata/graduation//FlumeData.1752737687931.tmp
```

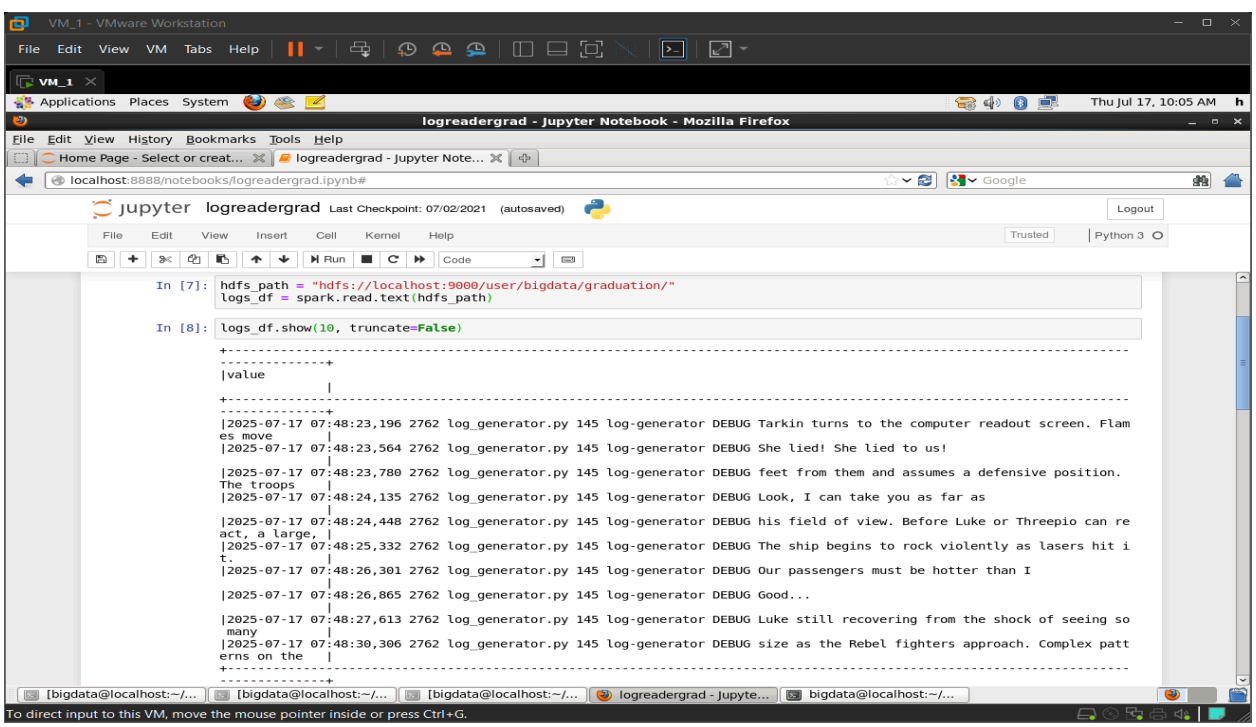


```
##! Define names for each Flume component
agent.sources = log-source           # Naming the source as "log-source"
agent.sinks = hdfs-sink              # Naming the sink as "hdfs-sink"
agent.channels = memory-channel      # Naming the channel (buffer between source and sink) as "memory-channel"
##! Source Configuration: Read from local log file using tail -F
agent.sources.log-source.type = exec
agent.sources.log-source.command = tail -F /home/bigdata/Desktop/logGenerator/logs/sample.log
# Flume will continuously monitor the local sample.log file using tail -F
# Every new line appended to sample.log will be captured by Flume in real-time
# Channel Configuration: Use in-memory buffer
agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000                # Maximum number of events the channel can store in memory
agent.channels.memory-channel.transactionCapacity = 100      # Maximum number of events to process in a single transaction
##! Sink Configuration: Write events to HDFS
agent.sinks.hdfs-sink.type = hdfs
agent.sinks.hdfs-sink.hdfs.path = hdfs://localhost:9000/user/bigdata/graduation/
##! Destination path in HDFS where log data will be saved
agent.sinks.hdfs-sink.hdfs.fileType = DataStream
##! DataStream means events are written to HDFS directly without large file buffering
agent.sinks.hdfs-sink.hdfs.writeFormat = Text
##! Events will be written as plain text (one log line per line in HDFS)
agent.sinks.hdfs-sink.hdfs.rollInterval = 10
# A new HDFS file is created every 10 seconds (even if the file is small)
##! Connect Source to Channel
agent.sources.log-source.channels = memory-channel
##! Connect Sink to Channel
agent.sinks.hdfs-sink.channel = memory-channel
```

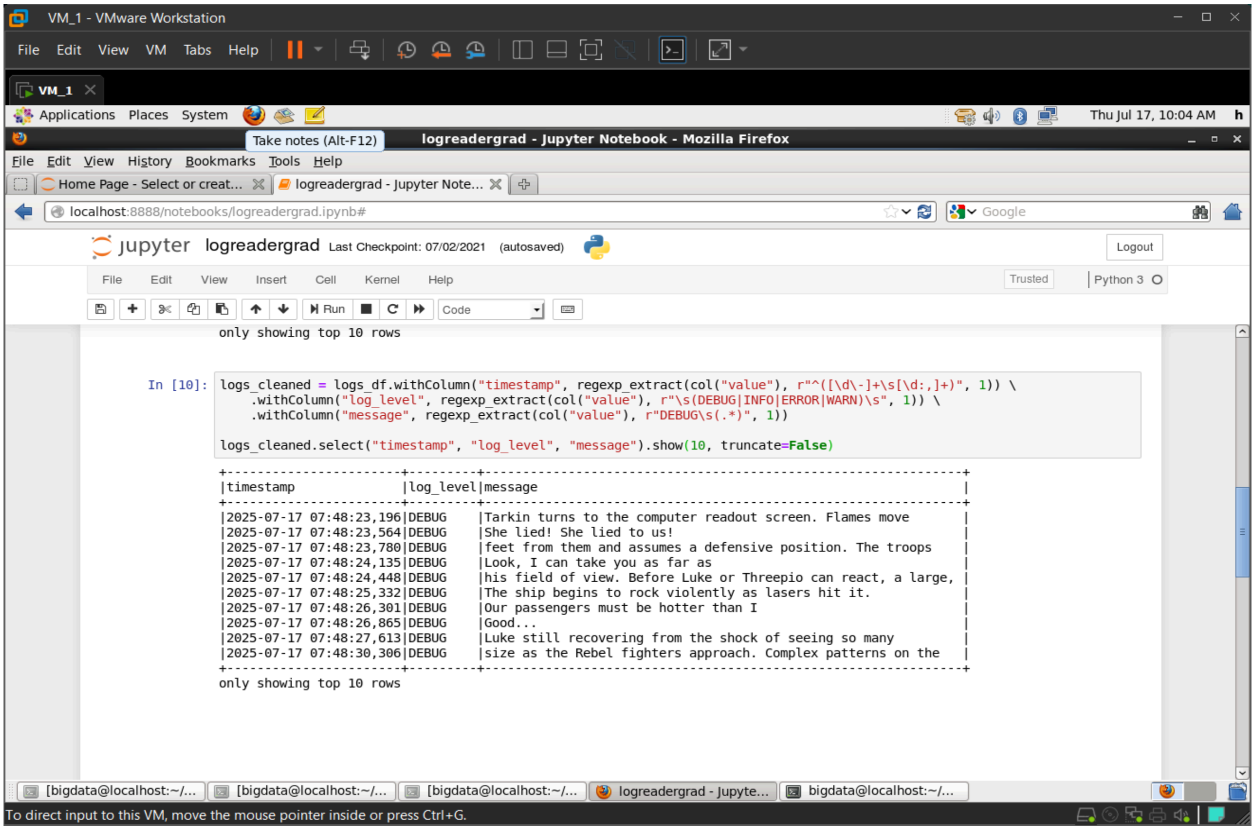
After that, I launched the Jupyter-Notebook and initialized PySpark.
[jupyter notebook]



- -



I performed some operations to improve the data presentation.



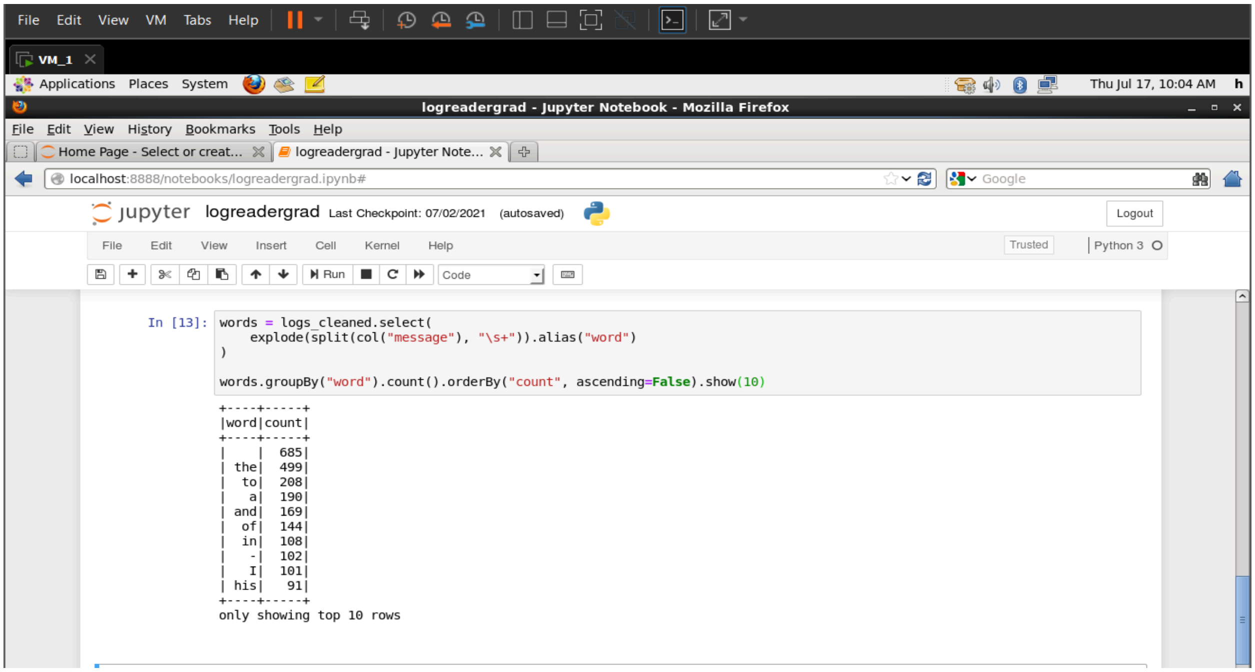
```
only showing top 10 rows
```

```
In [10]: logs_cleaned = logs_df.withColumn("timestamp", regexp_extract(col("value"), r"^([\d\-]+\s[\d:,]+)", 1)) \
             .withColumn("log_level", regexp_extract(col("value"), r"\s(DEBUG|INFO|ERROR|WARN)\s", 1)) \
             .withColumn("message", regexp_extract(col("value"), r"DEBUG\s(.*)", 1))

         logs_cleaned.select("timestamp", "log_level", "message").show(10, truncate=False)
```

```
+-----------------------+---------+-----------------------------------------------------------------+
|timestamp              |log_level|message                                                          |
+-----------------------+---------+-----------------------------------------------------------------+
|2025-07-17 07:48:23,196|DEBUG    |Tarkin turns to the computer readout screen. Flames move         |
|2025-07-17 07:48:23,564|DEBUG    |She lied! She lied to us!                                        |
|2025-07-17 07:48:23,780|DEBUG    |feet from them and assumes a defensive position. The troops      |
|2025-07-17 07:48:24,135|DEBUG    |Look, I can take you as far as                                   |
|2025-07-17 07:48:24,448|DEBUG    |his field of view. Before Luke or Threepio can react, a large,   |
|2025-07-17 07:48:25,332|DEBUG    |The ship begins to rock violently as lasers hit it.              |
|2025-07-17 07:48:26,301|DEBUG    |Our passengers must be hotter than I                             |
|2025-07-17 07:48:26,865|DEBUG    |Good...                                                          |
|2025-07-17 07:48:27,613|DEBUG    |Luke still recovering from the shock of seeing so many           |
|2025-07-17 07:48:30,306|DEBUG    |size as the Rebel fighters approach. Complex patterns on the     |
+-----------------------+---------+-----------------------------------------------------------------+
only showing top 10 rows
```

To direct input to this VM, move the mouse pointer inside or press Ctrl+G.

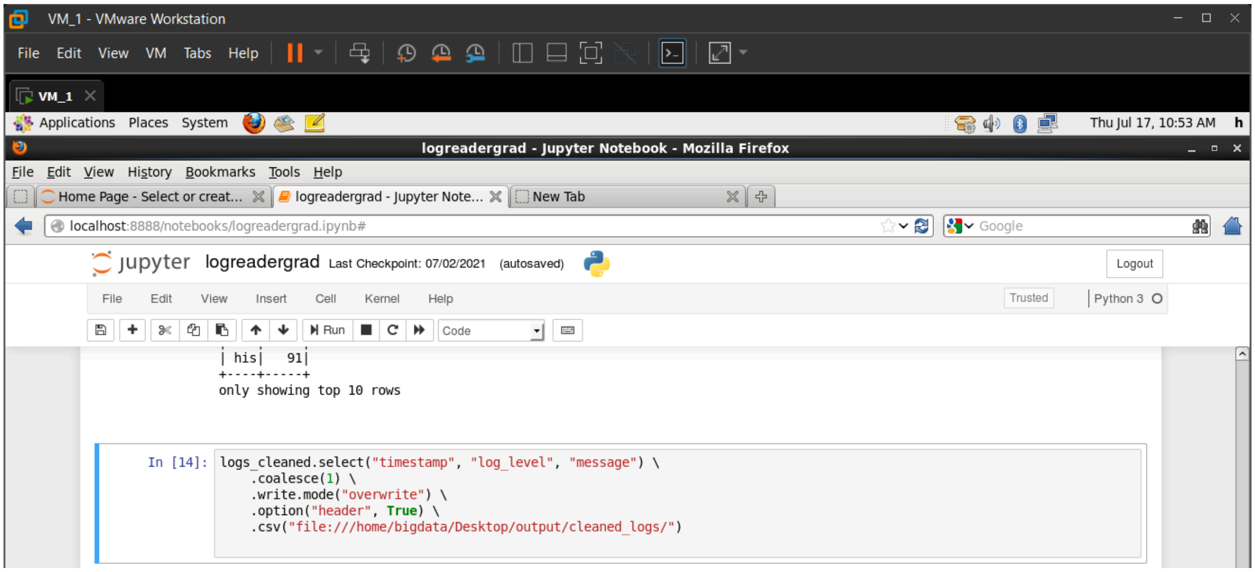


```
In [13]: words = logs_cleaned.select(
             explode(split(col("message"), "\s+")).alias("word")
         )

         words.groupBy("word").count().orderBy("count", ascending=False).show(10)
```

```
+----+-----+
|word|count|
+----+-----+
|    |  685|
| the|  499|
|  to|  208|
|   a|  190|
| and|  169|
|  of|  144|
|  in|  108|
|   -|  102|
|   I|  101|
| his|   91|
+----+-----+
only showing top 10 rows
```

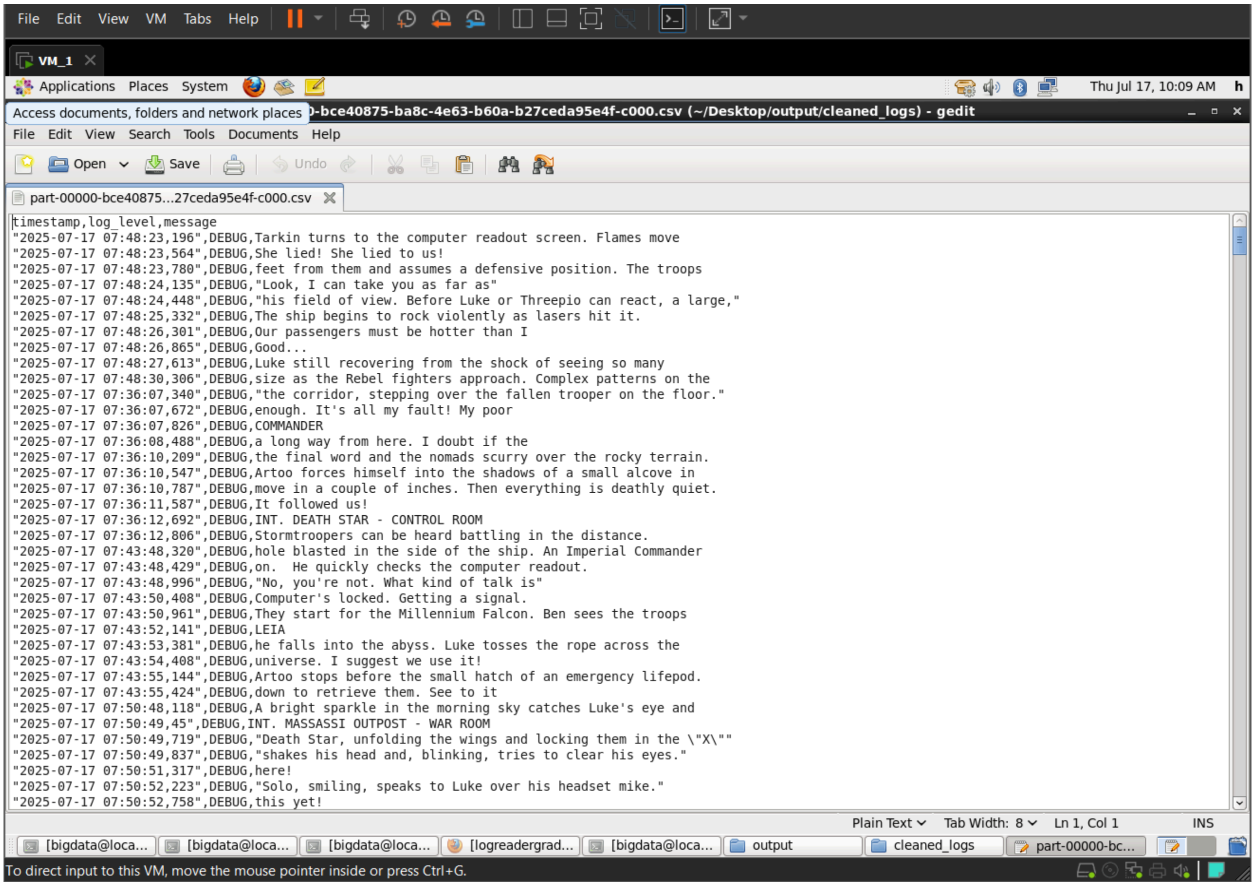
Finally, I saved the processed results using the following method:



```
| his|   91|
+----+-----+
only showing top 10 rows
```

```
In [14]: logs_cleaned.select("timestamp", "log_level", "message") \
             .coalesce(1) \
             .write.mode("overwrite") \
             .option("header", True) \
             .csv("file:///home/bigdata/Desktop/output/cleaned_logs/")
```

The .csv:



```
timestamp,log_level,message
"2025-07-17 07:48:23,196",DEBUG,Tarkin turns to the computer readout screen. Flames move
"2025-07-17 07:48:23,564",DEBUG,She lied! She lied to us!
"2025-07-17 07:48:23,780",DEBUG,feet from them and assumes a defensive position. The troops
"2025-07-17 07:48:24,135",DEBUG,"Look, I can take you as far as"
"2025-07-17 07:48:24,448",DEBUG,"his field of view. Before Luke or Threepio can react, a large,"
"2025-07-17 07:48:25,332",DEBUG,The ship begins to rock violently as lasers hit it.
"2025-07-17 07:48:26,301",DEBUG,Our passengers must be hotter than I
"2025-07-17 07:48:26,865",DEBUG,Good...
"2025-07-17 07:48:27,613",DEBUG,Luke still recovering from the shock of seeing so many
"2025-07-17 07:48:30,306",DEBUG,size as the Rebel fighters approach. Complex patterns on the
"2025-07-17 07:36:07,340",DEBUG,"the corridor, stepping over the fallen trooper on the floor."
"2025-07-17 07:36:07,672",DEBUG,enough. It's all my fault! My poor
"2025-07-17 07:36:07,826",DEBUG,COMMANDER
"2025-07-17 07:36:08,488",DEBUG,a long way from here. I doubt if the
"2025-07-17 07:36:10,209",DEBUG,the final word and the nomads scurry over the rocky terrain.
"2025-07-17 07:36:10,547",DEBUG,Artoo forces himself into the shadows of a small alcove in
"2025-07-17 07:36:10,787",DEBUG,move in a couple of inches. Then everything is deathly quiet.
"2025-07-17 07:36:11,587",DEBUG,It followed us!
"2025-07-17 07:36:12,692",DEBUG,INT. DEATH STAR - CONTROL ROOM
"2025-07-17 07:36:12,806",DEBUG,Stormtroopers can be heard battling in the distance.
"2025-07-17 07:43:48,320",DEBUG,hole blasted in the side of the ship. An Imperial Commander
"2025-07-17 07:43:48,429",DEBUG,on.  He quickly checks the computer readout.
"2025-07-17 07:43:48,996",DEBUG,"No, you're not. What kind of talk is"
"2025-07-17 07:43:50,408",DEBUG,Computer's locked. Getting a signal.
"2025-07-17 07:43:50,961",DEBUG,They start for the Millennium Falcon. Ben sees the troops
"2025-07-17 07:43:52,141",DEBUG,LEIA
"2025-07-17 07:43:53,381",DEBUG,he falls into the abyss. Luke tosses the rope across the
"2025-07-17 07:43:54,408",DEBUG,universe. I suggest we use it!
"2025-07-17 07:43:55,144",DEBUG,Artoo stops before the small hatch of an emergency lifepod.
"2025-07-17 07:43:55,424",DEBUG,down to retrieve them. See to it
"2025-07-17 07:50:48,118",DEBUG,A bright sparkle in the morning sky catches Luke's eye and
"2025-07-17 07:50:49,45",DEBUG,INT. MASSASSI OUTPOST - WAR ROOM
"2025-07-17 07:50:49,719",DEBUG,"Death Star, unfolding the wings and locking them in the \"X\""
"2025-07-17 07:50:49,837",DEBUG,"shakes his head and, blinking, tries to clear his eyes."
"2025-07-17 07:50:51,317",DEBUG,here!
"2025-07-17 07:50:52,223",DEBUG,"Solo, smiling, speaks to Luke over his headset mike."
"2025-07-17 07:50:52,758",DEBUG,this yet!
```

Notes:

- All directories used during the project, including log files and output folders, were created and managed locally on my **Desktop** for easier access and testing.

- All **scripts**, **Python code, configuration** files (including Flume configurations), and notebooks used in this project are fully available on my personal GitHub repository for reference and future use

  GitHub Profile: *https://github.com/khattabx*

  Repo Name: *NTI-Graduation-Project*