Presentation 3 / 29

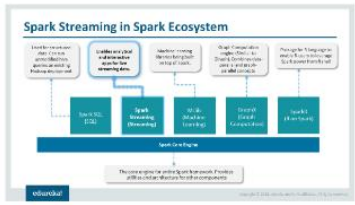
COURSE OUTLINE

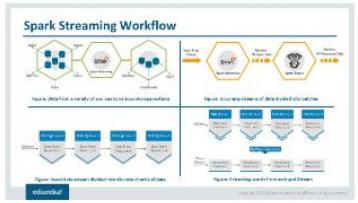
MODULE 11

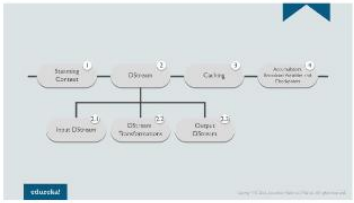
Introduction to Big Data Hadoop and Spark
Introduction to Python for Apache Spark
Functions, OOPs, Modules in Python
Deep Dive into Apache Spark Framework
Playing with Spark RDDs
Data Frames and Spark SQL
Machine Learning using Spark MLlib
Deep Dive into Spark MLlib
Understanding Apache Kafka and Apache Flume
Apache Spark Streaming - Processing Multiple Batches
Apache Spark Streaming - Data Sources
Implementing an End-to-End Project

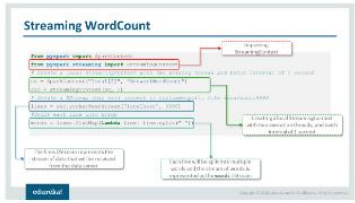
Presentation 4 / 29

Recap









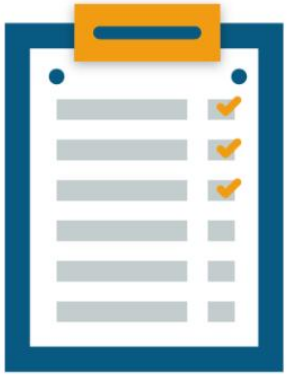
edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 5 / 29

Topics

- Introduction to Data Sources
- Streaming Data Sources
- Static Data Sources



edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 6 / 29

Objectives

After completing this module, you should be able to:

- Learn different streaming data sources such as Kafka, Twitter and Flume

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 7 / 29

Introduction to Data Sources

The Spark engine has the ability to access data from a variety of sources, including HDFS, Amazon S3, Kafka, Flume, Third Party API's. This data can be received as streaming input or can be read from a stored location. We can further utilize this data to perform analysis and other relevant operations.

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 8 / 29

Types of Data Sources

You can also categorize the various kind of data sources based on the kind of data they push into Spark engine. Data Sources can be majorly divided in two categories:

Streaming Data Sources that continuously push data inside the Spark Engine.

Streaming Data Sources

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 9 / 29

Types of Data Sources

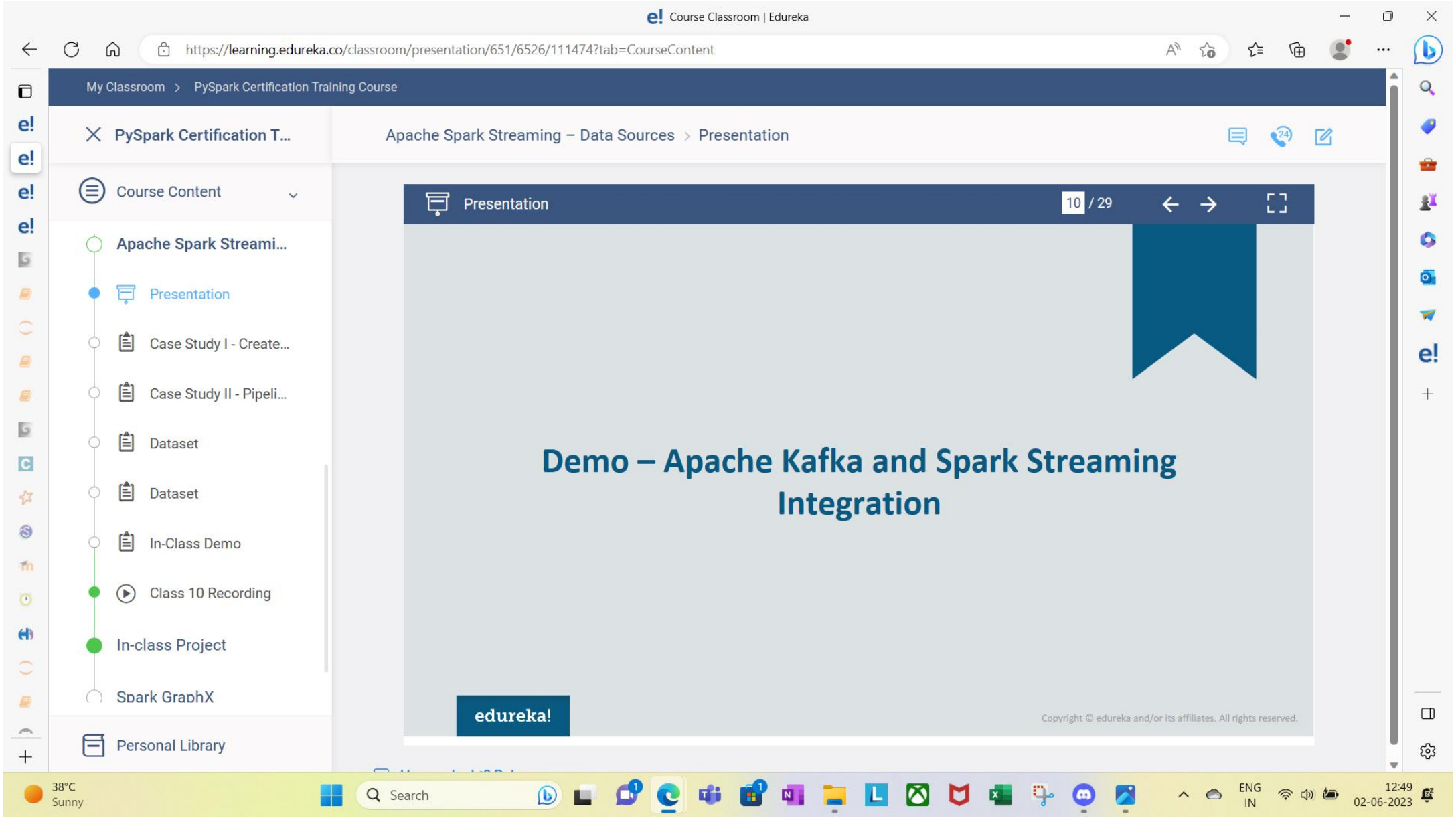
You can also categorize the various kind of data sources based on the kind of data they push into Spark engine. Data Sources can be majorly divided in two categories:

Static Data Sources

Static Data Sources that store data and Spark Engine fetches from this data when needed.

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
10 / 29
Demo – Apache Kafka and Spark Streaming Integration
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:49
02-06-2023

Presentation 11 / 29

Apache Kafka and Spark Streaming Integration

- Spark's support for Kafka closely aligns with the Spark Streaming project
- Kafka's performance and durability make it a platform that is well suited to servicing Spark Streaming processes
- Common usage scenarios for Kafka and Spark include Spark Streaming processes reading data from a Kafka topic and performing event processing on the data stream or a Spark process serving as a producer and writing output to a Kafka topic

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 12 / 29

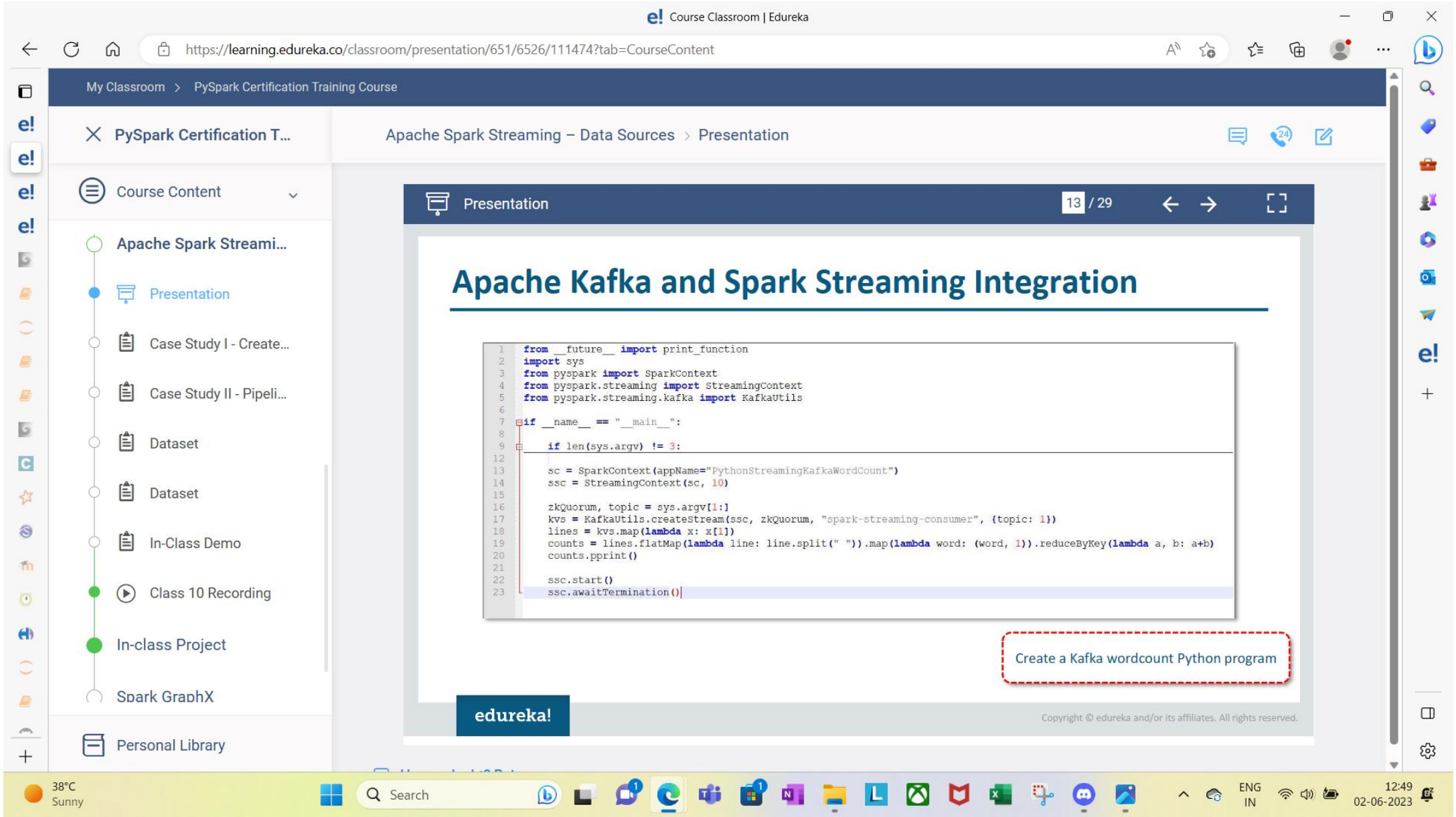
Apache Kafka and Spark Streaming Integration

```
/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/sentry-provider-common-1.5.1-cdh5.11.0.jar:/opt/cloudera/parcels/KAFKA-
3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/sentry-provider-db-1.5.1-cdh5.11.0.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/k
afka/bin/../libs/sentry-provider-file-1.5.1-cdh5.11.0.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/servlet-api-2
.5.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/shiro-core-1.2.3.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p
0.40/bin/../lib/kafka/bin/../libs/slf4j-api-1.7.25.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/slf4j-api-1.7.5.
jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/slf4j-log4j12-1.7.25.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.
p0.40/bin/../lib/kafka/bin/../libs/slf4j-log4j12-1.7.5.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/snappy-java-
1.1.2.6.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/stax-api-1.0-2.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.
0.p0.40/bin/../lib/kafka/bin/../libs/validation-api-1.1.0.Final.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/xml
enc-0.52.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/xz-1.0.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40
/bin/../lib/kafka/bin/../libs/zkclient-0.10.jar:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/zookeeper-3.4.10.jar:/o
pt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/bin/../lib/kafka/bin/../libs/zookeeper-3.4.5-cdh5.13.0.jar
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib:/usr/lib
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:os.name=Linux
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:os.arch=amd64
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:os.version=3.10.0-693.5.2.el7.x86_64
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:user.name=edureka_321047
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:user.home=/mnt/home/edureka_321047
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Client environment:user.dir=/mnt/home/edureka_321047
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Initiating client connection, connectString=ip-20-0-21-161.ec2.internal:2181 sessionTimeout=30000 watche
r=org.I0Itec.zkclient.ZkClient@59fd97a8
18/07/21 07:34:30 INFO zkclient.ZkClient: Waiting for keeper state SyncConnected
18/07/21 07:34:30 INFO zookeeper.ClientCnxn: Opening socket connection to server ip-20-0-21-161.ec2.internal/20.0.21.161:2181. Will not attempt to a
uthenticate using SASL (unknown error)
18/07/21 07:34:30 INFO zookeeper.ClientCnxn: Socket connection established to ip-20-0-21-161.ec2.internal/20.0.21.161:2181, initiating session
18/07/21 07:34:30 INFO zookeeper.ClientCnxn: Session establishment complete on server ip-20-0-21-161.ec2.internal/20.0.21.161:2181, sessionid = 0x76
3eeae585c624f, negotiated timeout = 30000
18/07/21 07:34:30 INFO zkclient.ZkClient: zookeeper state changed (SyncConnected)
18/07/21 07:34:30 INFO admin.AdminUtils$: Topic creation {"version":1,"partitions":{"0":[296]}}
Created topic "wordcounttopic".
18/07/21 07:34:30 INFO zkclient.ZkEventThread: Terminate ZkClient event thread.
18/07/21 07:34:30 INFO zookeeper.ZooKeeper: Session: 0x763eeae585c624f closed
18/07/21 07:34:30 INFO zookeeper.ClientCnxn: EventThread shut down for session: 0x763eeae585c624f
[edureka_321047@ip-20-0-41-190 ~]$
```

Creating a Kafka topic named "wordcounttopic"

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
13 / 29
Apache Kafka and Spark Streaming Integration
from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from pyspark.streaming.kafka import KafkaUtils
if __name__ == "__main__":
if len(sys.argv) != 3:
sc = SparkContext(appName="PythonStreamingKafkaWordCount")
ssc = StreamingContext(sc, 10)
zkQuorum, topic = sys.argv[1:]
kvs = KafkaUtils.createStream(ssc, zkQuorum, "spark-streaming-consumer", {topic: 1})
lines = kvs.map(lambda x: x[1])
counts = lines.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a+b)
counts.pprint()
ssc.start()
ssc.awaitTermination()
Create a Kafka wordcount Python program
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:49
02-06-2023

Presentation 14 / 29

Apache Kafka and Spark Streaming Integration

```
[edureka_321047@ip-20-0-41-190 ~]$ spark2-submit --master yarn kafka_wordcount.py ip-20-0-21-161.ec2.internal:2181 wordcounttopic
18/07/21 09:46:20 INFO spark.SparkContext: Running Spark version 2.1.0.cloudera2
18/07/21 09:46:21 INFO spark.SecurityManager: Changing view acls to: edureka_321047
18/07/21 09:46:21 INFO spark.SecurityManager: Changing modify acls to: edureka_321047
18/07/21 09:46:21 INFO spark.SecurityManager: Changing view acls groups to:
18/07/21 09:46:21 INFO spark.SecurityManager: Changing modify acls groups to:
18/07/21 09:46:21 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(edureka_
321047); groups with view permissions: Set(); users  with modify permissions: Set(edureka_321047); groups with modify permissions: Set()
18/07/21 09:46:22 INFO util.Utils: Successfully started service 'sparkDriver' on port 38294.
18/07/21 09:46:22 INFO spark.SparkEnv: Registering MapOutputTracker
18/07/21 09:46:22 INFO spark.SparkEnv: Registering BlockManagerMaster
18/07/21 09:46:22 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
18/07/21 09:46:22 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/07/21 09:46:22 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-e3ba80c7-15df-4942-be4e-b2fff95c3b3c
18/07/21 09:46:22 INFO memory.MemoryStore: MemoryStore started with capacity 93.3 MB
18/07/21 09:46:22 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/07/21 09:46:23 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers
18/07/21 09:46:23 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (4096 MB per
container)
18/07/21 09:46:23 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
18/07/21 09:46:23 INFO yarn.Client: Setting up container launch context for our AM
18/07/21 09:46:23 INFO yarn.Client: Setting up the launch environment for our AM container
18/07/21 09:46:23 INFO yarn.Client: Preparing resources for our AM container
18/07/21 09:46:25 INFO yarn.Client: Uploading resource file:/tmp/spark-2116b936-690d-4a33-8db0-e9a1cc921637/__spark_conf__3308945078087702336.zip ->
 hdfs://nameservice1/user/edureka_321047/.sparkStaging/application_1528714825862_16231/__spark_conf__.zip
18/07/21 09:46:25 INFO spark.SecurityManager: Changing view acls to: edureka_321047
18/07/21 09:46:25 INFO spark.SecurityManager: Changing modify acls to: edureka_321047
18/07/21 09:46:25 INFO spark.SecurityManager: Changing view acls groups to:
18/07/21 09:46:25 INFO spark.SecurityManager: Changing modify acls groups to:
```

Running the Kafka wordcount program on the shell

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 15 / 29

Apache Kafka and Spark Streaming Integration

```
ip-20-0-41-190 login: edureka_321047
Password:
Last login: Sat Jul 21 09:44:10 on pts/33
[edureka_321047@ip-20-0-41-190 ~]$ kafka-console-producer --broker-list ip-20-0-31-221.ec2.internal:9092 --topic wordcounttopic
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/lib/kafka/libs/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLogg
erBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/KAFKA-3.0.0-1.3.0.0.p0.40/lib/kafka/libs/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLogge
rBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/07/21 09:48:48 INFO producer.ProducerConfig: ProducerConfig values:
        acks = 1
        batch.size = 16384
        bootstrap.servers = [ip-20-0-31-221.ec2.internal:9092]
        buffer.memory = 33554432
        client.id = console-producer
        compression.type = none
        connections.max.idle.ms = 540000
        enable.idempotence = false
        interceptor.classes = null
        key.serializer = class org.apache.kafka.common.serialization.ByteArraySerializer
        linger.ms = 1000
        max.block.ms = 60000
        max.in.flight.requests.per.connection = 5
        max.request.size = 1048576
        metadata.max.age.ms = 300000
        metric.reporters = []
        metrics.num.samples = 2
        metrics.recording.level = INFO
        metrics.sample.window.ms = 30000
        partitioner.class = class org.apache.kafka.clients.producer.internals.DefaultPartitioner
        receive.buffer.bytes = 32768
        reconnect.backoff.max.ms = 1000
        reconnect.backoff.ms = 50
        request.timeout.ms = 1500
```

Running the Kafka Producer on shell for pushing the data to the topic

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 16 / 29

Apache Kafka and Spark Streaming Integration

```
        ssl.provider = null
        ssl.secure.random.implementation = null
        ssl.trustmanager.algorithm = PKIX
        ssl.truststore.location = null
        ssl.truststore.password = null
        ssl.truststore.type = JKS
        transaction.timeout.ms = 60000
        transactional.id = null
        value.serializer = class org.apache.kafka.common.serialization.ByteArraySerializer

18/07/21 09:48:48 INFO utils.AppInfoParser: Kafka version : 0.11.0-kafka-3.0.0
18/07/21 09:48:48 INFO utils.AppInfoParser: Kafka commitId : unknown
>Hello
>Hello
>Hello
>Hello
>Hello
>Edureka
>Edurek
>Hello
>Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello Spark
>Hello SparkHello Spark
>Hello Spark
>Hello Spark
>
```

Here, we are pushing the given messages to the topic

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 17 / 29

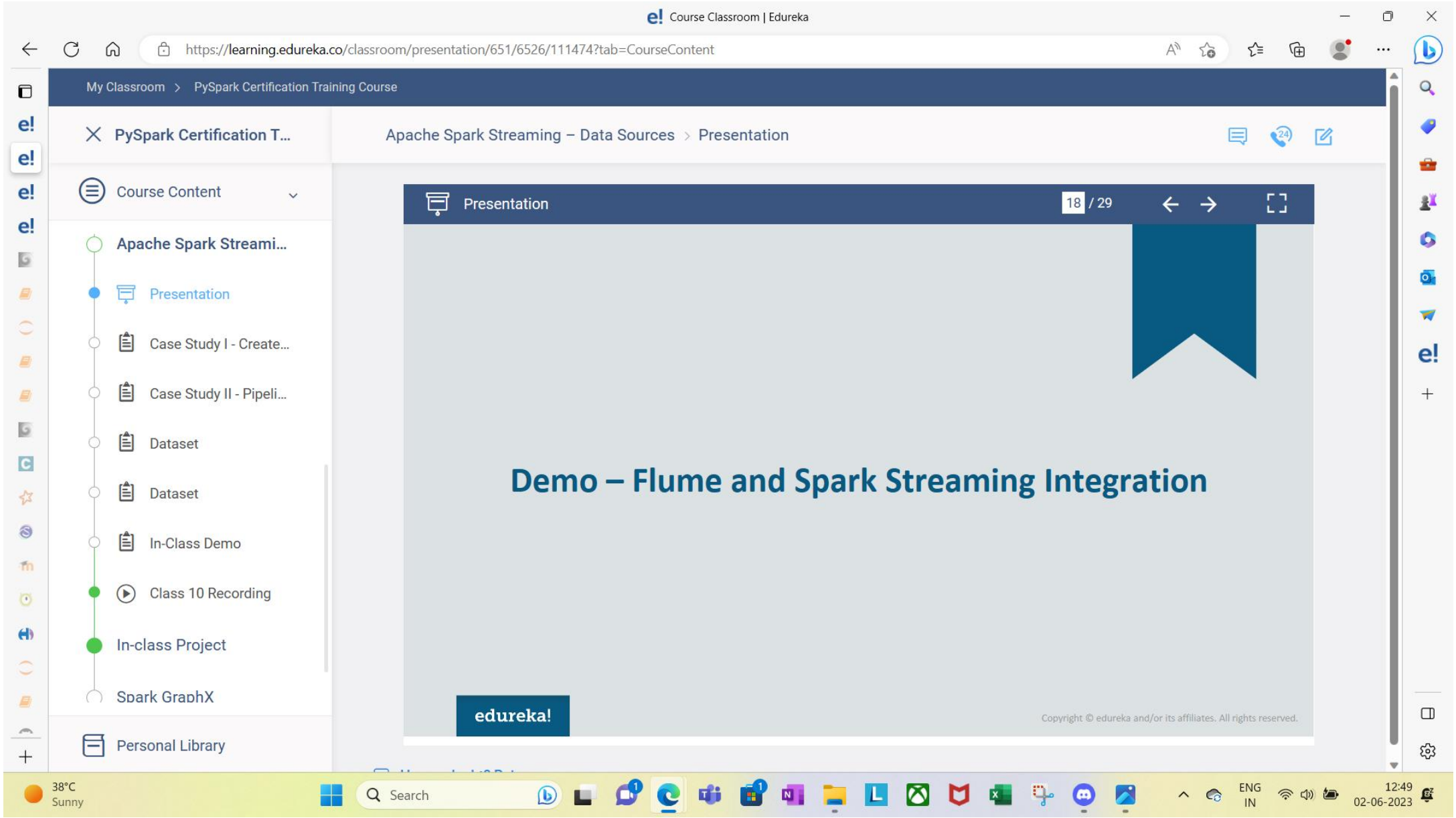
Apache Kafka and Spark Streaming Integration

```
18/07/21 09:51:10 INFO storage.BlockManagerInfo: Added broadcast_62_piece0 in memory on ip-20-0-31-4.ec2.internal:36951 (size: 4.1 KB, free: 366.3 M
B)
18/07/21 09:51:10 INFO spark.MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 28 to 20.0.31.4:33804
18/07/21 09:51:10 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 28 is 174 bytes
18/07/21 09:51:10 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 112.0 (TID 143) in 91 ms on ip-20-0-31-4.ec2.internal (executor 1) (1/1)
18/07/21 09:51:10 INFO cluster.YarnScheduler: Removed TaskSet 112.0, whose tasks have all completed, from pool
18/07/21 09:51:10 INFO scheduler.DAGScheduler: ResultStage 112 (runJob at PythonRDD.scala:441) finished in 0.091 s
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Job 56 finished: runJob at PythonRDD.scala:441, took 0.293799 s
18/07/21 09:51:10 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:441
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Got job 57 (runJob at PythonRDD.scala:441) with 1 output partitions
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Final stage: ResultStage 114 (runJob at PythonRDD.scala:441)
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 113)
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Missing parents: List()
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Submitting ResultStage 114 (PythonRDD[227] at RDD at PythonRDD.scala:48), which has no missing parent
s
18/07/21 09:51:10 INFO memory.MemoryStore: Block broadcast_63 stored as values in memory (estimated size 7.2 KB, free 93.1 MB)
18/07/21 09:51:10 INFO memory.MemoryStore: Block broadcast_63_piece0 stored as bytes in memory (estimated size 4.1 KB, free 93.1 MB)
18/07/21 09:51:10 INFO storage.BlockManagerInfo: Added broadcast_63_piece0 in memory on 20.0.41.190:42344 (size: 4.1 KB, free: 93.2 MB)
18/07/21 09:51:10 INFO spark.SparkContext: Created broadcast 63 from broadcast at DAGScheduler.scala:991
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 114 (PythonRDD[227] at RDD at PythonRDD.scala:48)
18/07/21 09:51:10 INFO cluster.YarnScheduler: Adding task set 114.0 with 1 tasks
18/07/21 09:51:10 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 114.0 (TID 144, ip-20-0-31-4.ec2.internal, executor 1, partition 1, NODE
_LOCAL, 5678 bytes)
18/07/21 09:51:10 INFO storage.BlockManagerInfo: Added broadcast_63_piece0 in memory on ip-20-0-31-4.ec2.internal:36951 (size: 4.1 KB, free: 366.3 M
B)
18/07/21 09:51:10 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 114.0 (TID 144) in 150 ms on ip-20-0-31-4.ec2.internal (executor 1) (1/1
)
18/07/21 09:51:10 INFO cluster.YarnScheduler: Removed TaskSet 114.0, whose tasks have all completed, from pool
18/07/21 09:51:10 INFO scheduler.DAGScheduler: ResultStage 114 (runJob at PythonRDD.scala:441) finished in 0.151 s
18/07/21 09:51:10 INFO scheduler.DAGScheduler: Job 57 finished: runJob at PythonRDD.scala:441, took 0.155179 s
-------------------------------------------
Time: 2018-07-21 09:51:10
-------------------------------------------
(u'Spark', 5)
(u'SparkHello', 1)
(u'Hello', 5)
```

We can see, the Kafka wordcount program displaying each word along with it's frequency

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
18 / 29
Demo – Flume and Spark Streaming Integration
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:49
02-06-2023

Presentation 19 / 29

Flume and Spark Streaming Integration

We are working on a solution to monitor any application by analyzing the log files in Apache Spark through streaming module. We will build a flume and spark pipeline to parse the logs.

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 20 / 29

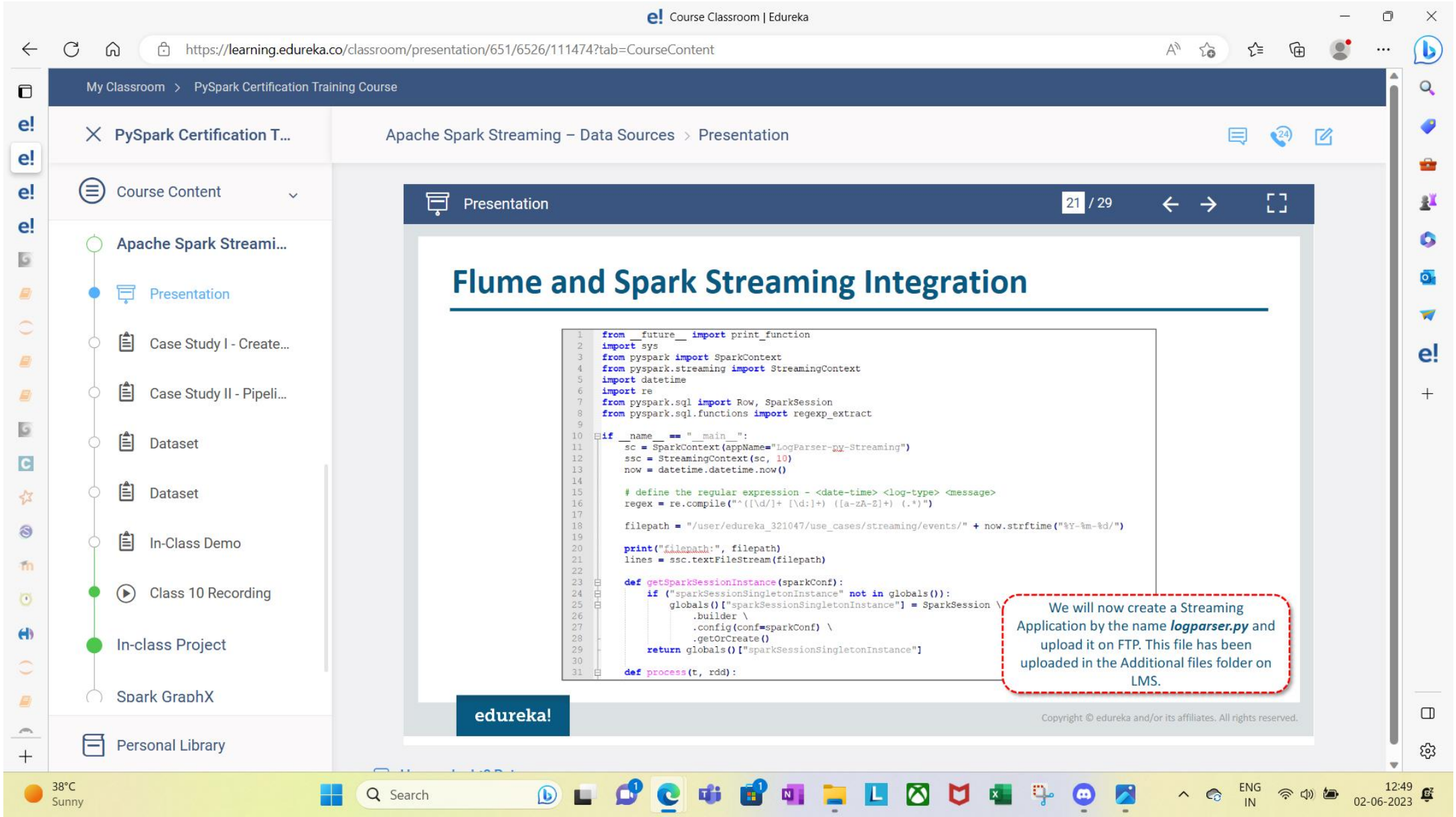
Flume and Spark Streaming Integration

```
# Name the components on this agent
stlog.sources = r1
stlog.sinks = k1
stlog.channels = c1

# Describe/configure the source
stlog.sources.r1.type = exec
stlog.sources.r1.command = tail -F -s 2 /mnt/home/edureka_321047/av/workspace/logs/event2.log

# Describe the sink
stlog.sinks.k1.type = hdfs
stlog.sinks.k1.hdfs.path = use_cases/streaming/events/%Y-%m-%d/
stlog.sinks.k1.hdfs.filePrefix = events-
stlog.sinks.k1.hdfs.fileSuffix = .log
stlog.sinks.k1.hdfs.useLocalTimeStamp = true
stlog.sinks.k1.hdfs.fileType = DataStream

# Use a channel stlogich buffers events in memory
stlog.channels.c1.type = memory
stlog.channels.c1.capacity = 1000
stlog.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
stlog.sources.r1.channels = c1
stlog.sinks.k1.channel = c1
```

We will first create a config file for flume by the name ***stlog.conf*** and upload it on FTP. This file has been uploaded in the Additional files folder on LMS.

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
21 / 29
Flume and Spark Streaming Integration
from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
import datetime
import re
from pyspark.sql import Row, SparkSession
from pyspark.sql.functions import regexp_extract
if __name__ == "__main__":
sc = SparkContext(appName="LogParser-py-Streaming")
ssc = StreamingContext(sc, 10)
now = datetime.datetime.now()
# define the regular expression - <date-time> <log-type> <message>
regex = re.compile("^([\d/]+ [\d:]+) ([a-zA-Z]+) (.*)")
filepath = "/user/edureka_321047/use_cases/streaming/events/" + now.strftime("%Y-%m-%d/")
print("filepath:", filepath)
lines = ssc.textFileStream(filepath)
def getSparkSessionInstance(sparkConf):
if ("sparkSessionSingletonInstance" not in globals()):
globals()["sparkSessionSingletonInstance"] = SparkSession \
.builder \
.config(conf=sparkConf) \
.getOrCreate()
return globals()["sparkSessionSingletonInstance"]
def process(t, rdd):
We will now create a Streaming Application by the name logparser.py and upload it on FTP. This file has been uploaded in the Additional files folder on LMS.
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C Sunny
Search
ENG IN
12:49 02-06-2023

Presentation 22 / 29

Flume and Spark Streaming Integration

```
[edureka_321047@ip-20-0-41-202 ~]$ flume-ng agent -n stlog -c conf -f stlog.conf -Dflume.root.logger=INFO,console
Info: Including Hadoop libraries found via (/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/bin/hbase) for HBASE access
Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed in a future release
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java/jdk1.8.0_144-cloudera/bin/java -Xmx20m -Dflume.root.logger=INFO,console -cp 'conf:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4
/lib/flume-ng/lib/*:/etc/hadoop/conf:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop/lib/*:/opt/cloudera/parcels/C
DH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop/.//*:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop-hd
fs/./:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop-hdfs/lib/*:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4
/lib/hadoop/libexec/../../hadoop-hdfs/.//*:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop-yarn/lib/*:/opt/clouder
a/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop-yarn/.//*:/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/lib/*:/opt/cloudera/p
arcels/CDH/lib/hadoop-mapreduce/.//*:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../conf:/usr/java/jdk1.8.0_144-cloudera/lib/too
ls.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/..:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/act
ivation-1.1.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/apacheds-i18n-2.0.0-M15.jar:/opt/cloudera/parcels/CDH-5.11.1-
1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/apacheds-kerberos-codec-2.0.0-M15.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/a
pi-asn1-api-1.0.0-M20.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/api-util-1.0.0-M20.jar:/opt/cloudera/parcels/CDH-5.
11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/asm-3.2.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/avro.jar:/opt/cloudera
/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/../lib/aws-java-sdk-core-1.10.6.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/b
in/../lib/aws-java-sdk-dynamodb-1.10.
```

Run this command on console to start the flume agent :

flume-ng agent -n stlog -c conf -f stlog.conf -Dflume.root.logger=INFO,console

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Presentation 23 / 29

Flume and Spark Streaming Integration

```
[edureka_321047@ip-20-0-41-202 ~]$ spark2-submit logparser_streaming.py --deploy-mode client > /mnt/home/edureka_321047/av/workspace/logs/streaming_
output.log 2>&1
```

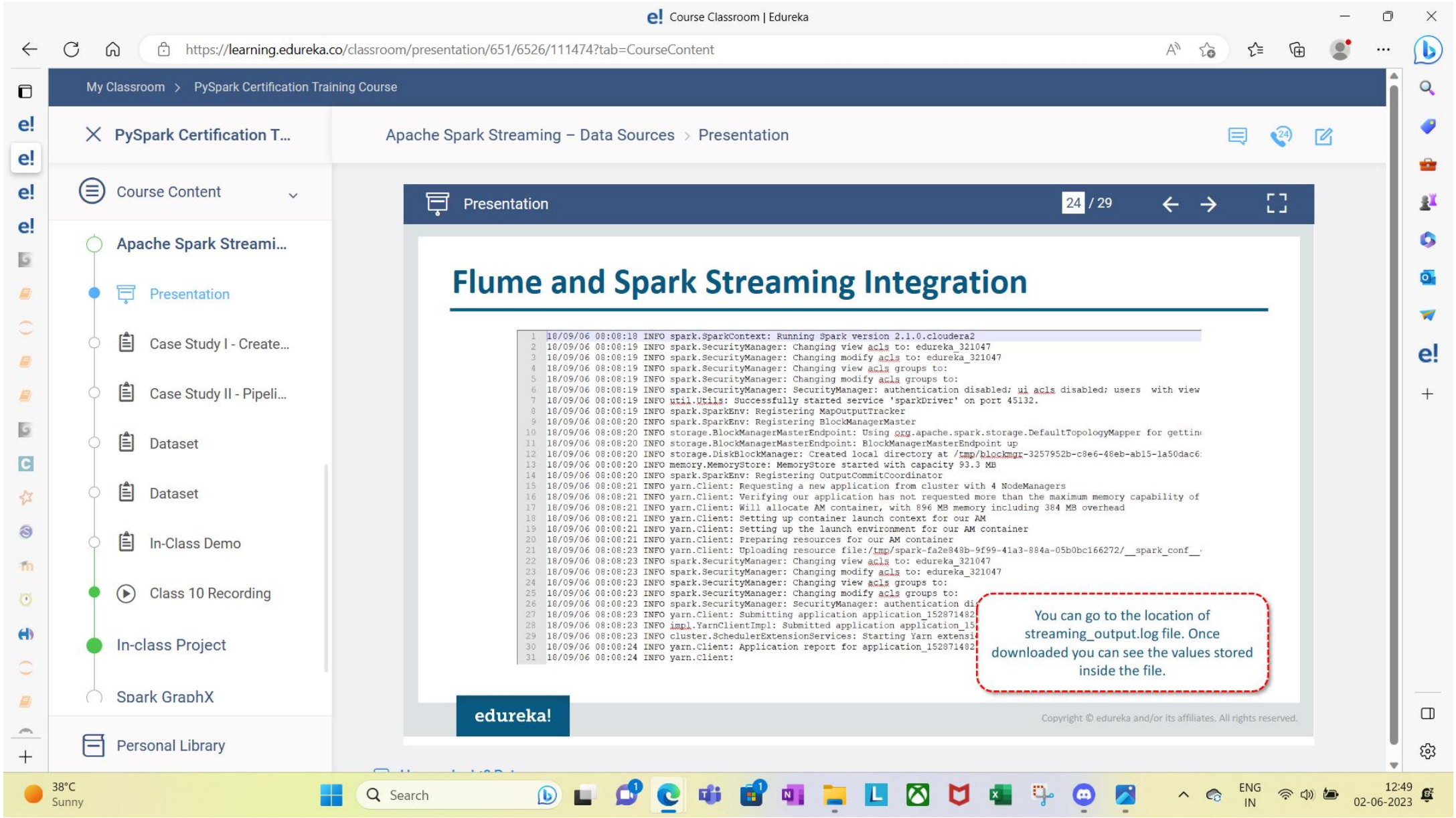
This command contains the path of the location where your streaming output will be stored. Here the name of the file is ***streaming_output.log***.

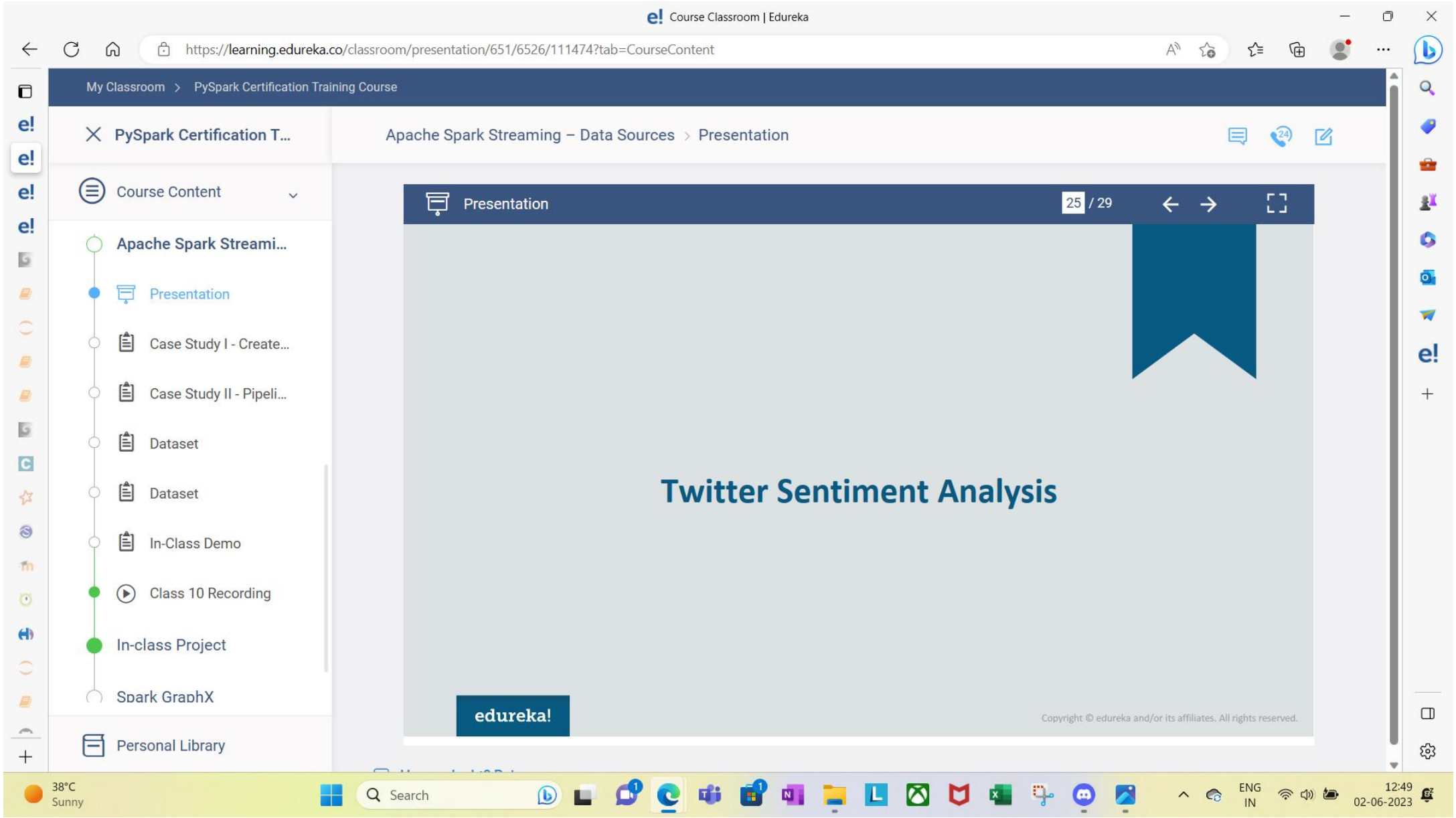
Run this command on console to start the flume agent :

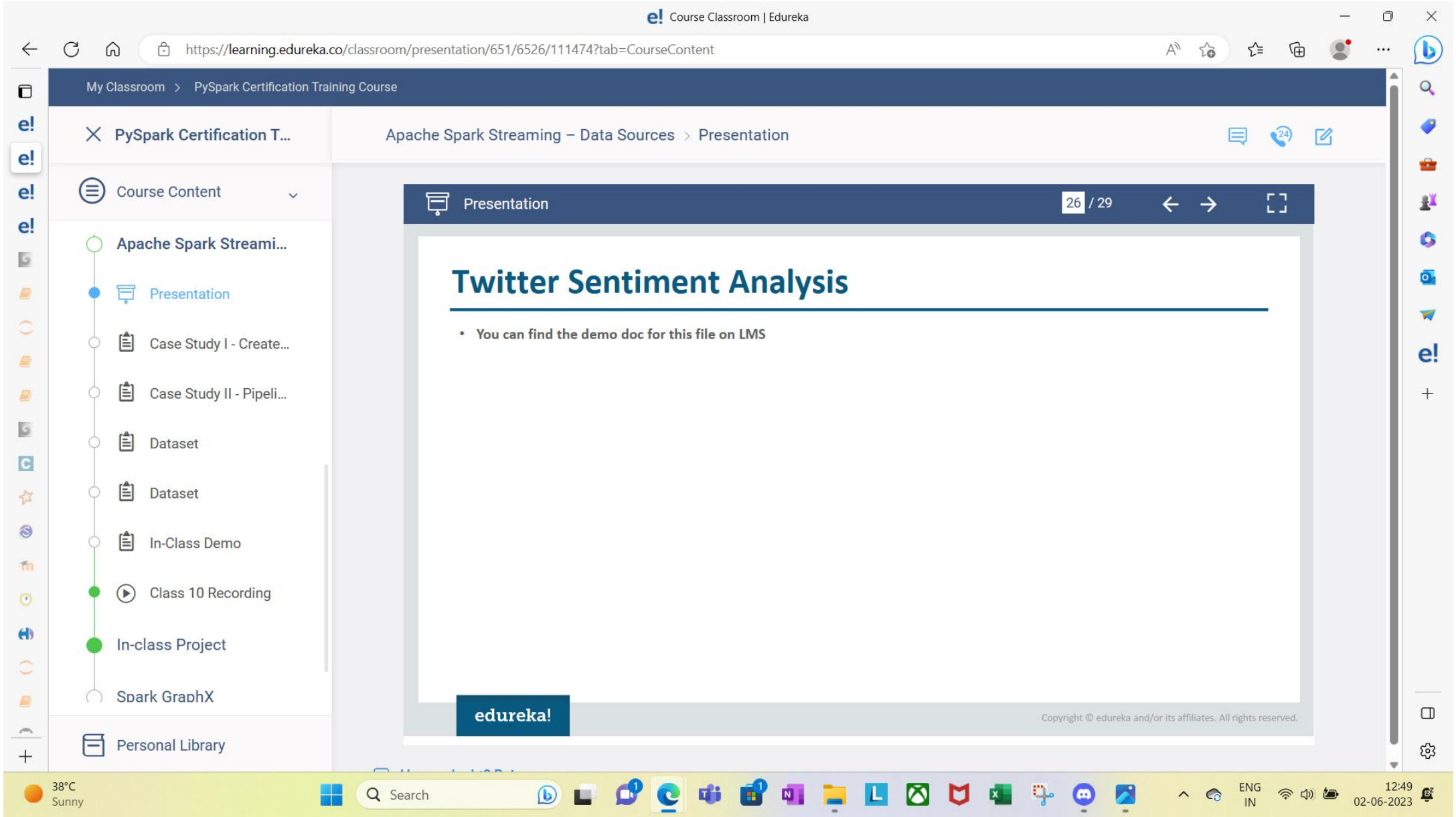
spark2-submit logparser_streaming.py --deploy-mode client > /mnt/home/edureka_321047/av/workspace/logs/streaming_output.log 2>&1

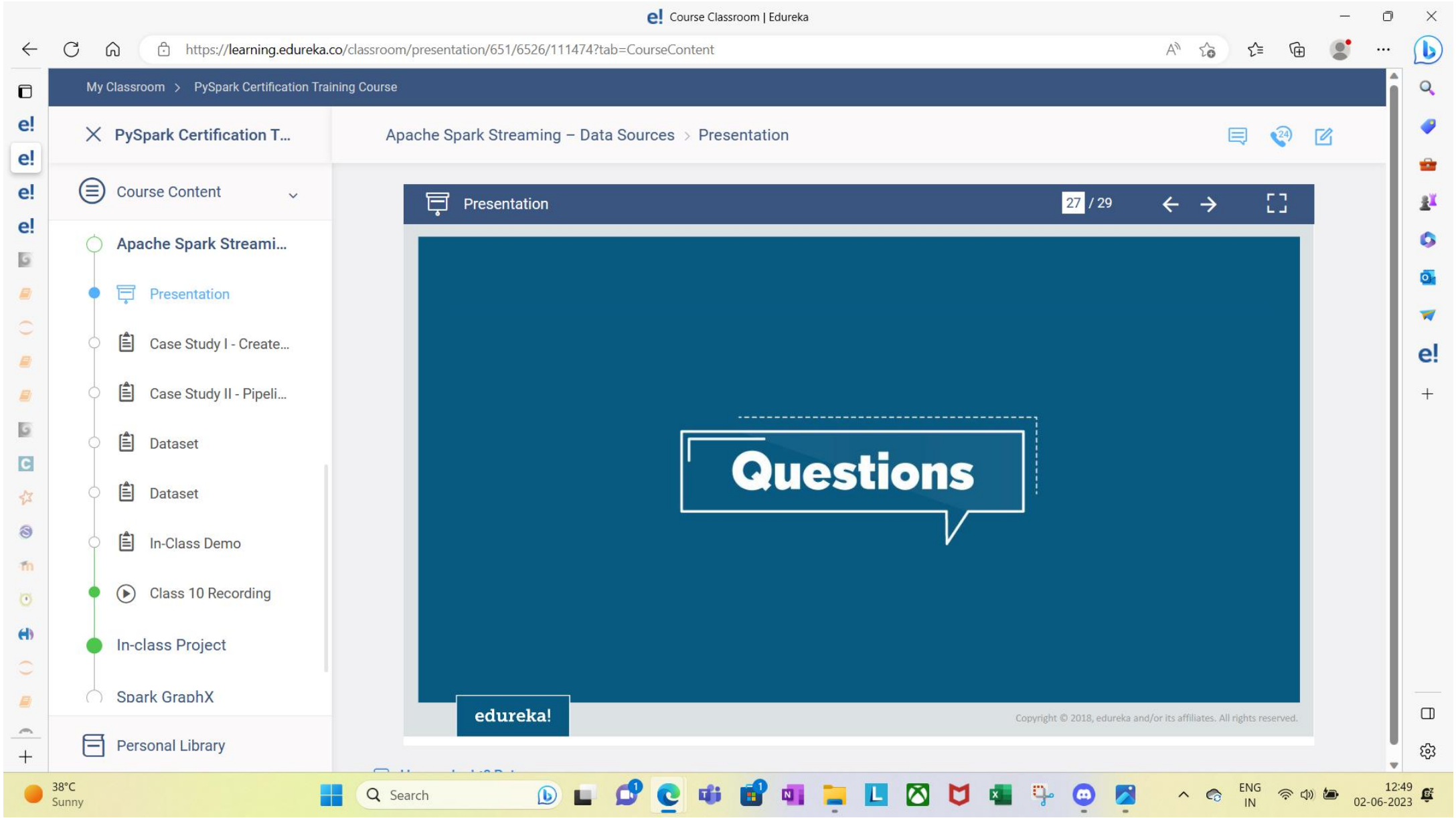
edureka!

Copyright © edureka and/or its affiliates. All rights reserved.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Apache Spark Streaming – Data Sources > Presentation
Presentation
24 / 29
Flume and Spark Streaming Integration
18/09/06 08:08:18 INFO spark.SparkContext: Running Spark version 2.1.0.cloudera2
18/09/06 08:08:19 INFO spark.SecurityManager: Changing view acls to: edureka_321047
18/09/06 08:08:19 INFO spark.SecurityManager: Changing modify acls to: edureka_321047
18/09/06 08:08:19 INFO spark.SecurityManager: Changing view acls groups to:
18/09/06 08:08:19 INFO spark.SecurityManager: Changing modify acls groups to:
18/09/06 08:08:19 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view
18/09/06 08:08:19 INFO util.Utils: Successfully started service 'sparkDriver' on port 45132.
18/09/06 08:08:19 INFO spark.SparkEnv: Registering MapOutputTracker
18/09/06 08:08:20 INFO spark.SparkEnv: Registering BlockManagerMaster
18/09/06 08:08:20 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for gettin
18/09/06 08:08:20 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/09/06 08:08:20 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-3257952b-c8e6-48eb-ab15-1a50dac6
18/09/06 08:08:20 INFO memory.MemoryStore: MemoryStore started with capacity 93.3 MB
18/09/06 08:08:20 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/09/06 08:08:21 INFO yarn.Client: Requesting a new application from cluster with 4 NodeManagers
18/09/06 08:08:21 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of
18/09/06 08:08:21 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
18/09/06 08:08:21 INFO yarn.Client: Setting up container launch context for our AM
18/09/06 08:08:21 INFO yarn.Client: Setting up the launch environment for our AM container
18/09/06 08:08:21 INFO yarn.Client: Preparing resources for our AM container
18/09/06 08:08:23 INFO yarn.Client: Uploading resource file:/tmp/spark-fa2e848b-9f99-41a3-884a-05b0bc166272/__spark_conf__
18/09/06 08:08:23 INFO spark.SecurityManager: Changing view acls to: edureka_321047
18/09/06 08:08:23 INFO spark.SecurityManager: Changing modify acls to: edureka_321047
18/09/06 08:08:23 INFO spark.SecurityManager: Changing view acls groups to:
18/09/06 08:08:23 INFO spark.SecurityManager: Changing modify acls groups to:
18/09/06 08:08:23 INFO spark.SecurityManager: SecurityManager: authentication di
18/09/06 08:08:23 INFO yarn.Client: Submitting application application_1528714822
18/09/06 08:08:23 INFO impl.YarnClientImpl: Submitted application application_15
18/09/06 08:08:23 INFO cluster.SchedulerExtensionServices: Starting Yarn extensi
18/09/06 08:08:24 INFO yarn.Client: Application report for application_1528714822
18/09/06 08:08:24 INFO yarn.Client:
You can go to the location of streaming_output.log file. Once downloaded you can see the values stored inside the file.
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C Sunny
Search
ENG IN
12:49 02-06-2023


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
25 / 29
Twitter Sentiment Analysis
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:49
02-06-2023


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
26 / 29
Twitter Sentiment Analysis
You can find the demo doc for this file on LMS
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:49
02-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6526/111474?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Apache Spark Streaming – Data Sources > Presentation
Course Content
Apache Spark Streami...
Presentation
Case Study I - Create...
Case Study II - Pipeli...
Dataset
Dataset
In-Class Demo
Class 10 Recording
In-class Project
Spark GraphX
Personal Library
Presentation
27 / 29
Questions
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:49
02-06-2023