Presentation 3 / 118

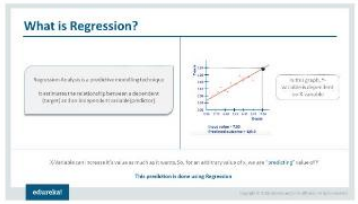
COURSE OUTLINE

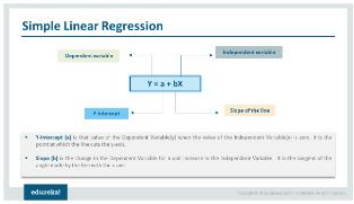
MODULE 09

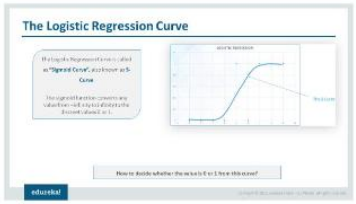
Introduction to Big Data Hadoop and Spark
Introduction to Python for Apache Spark
Functions, OOPs, Modules in Python
Deep Dive into Apache Spark Framework
Playing with Spark RDDs
Data Frames and Spark SQL
Machine Learning using Spark MLlib
Deep Dive into Spark MLlib
Understanding Apache Kafka and Apache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

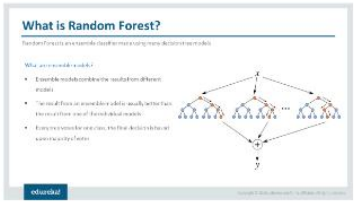
Presentation 4 / 118

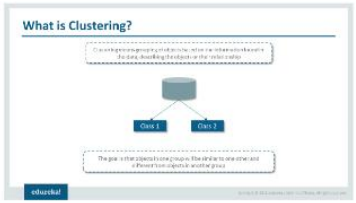
Recap









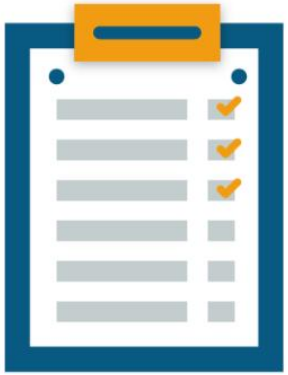


edureka!

Presentation 5 / 118

Topics

- Why is Kafka needed ?
- What is Kafka ?
- Kafka features and components
- Kafka Architecture
- Kafka Scaling and Cluster
- Flume
- Flume Architecture and Features
- Flume Dataflows and Configuration



edureka!

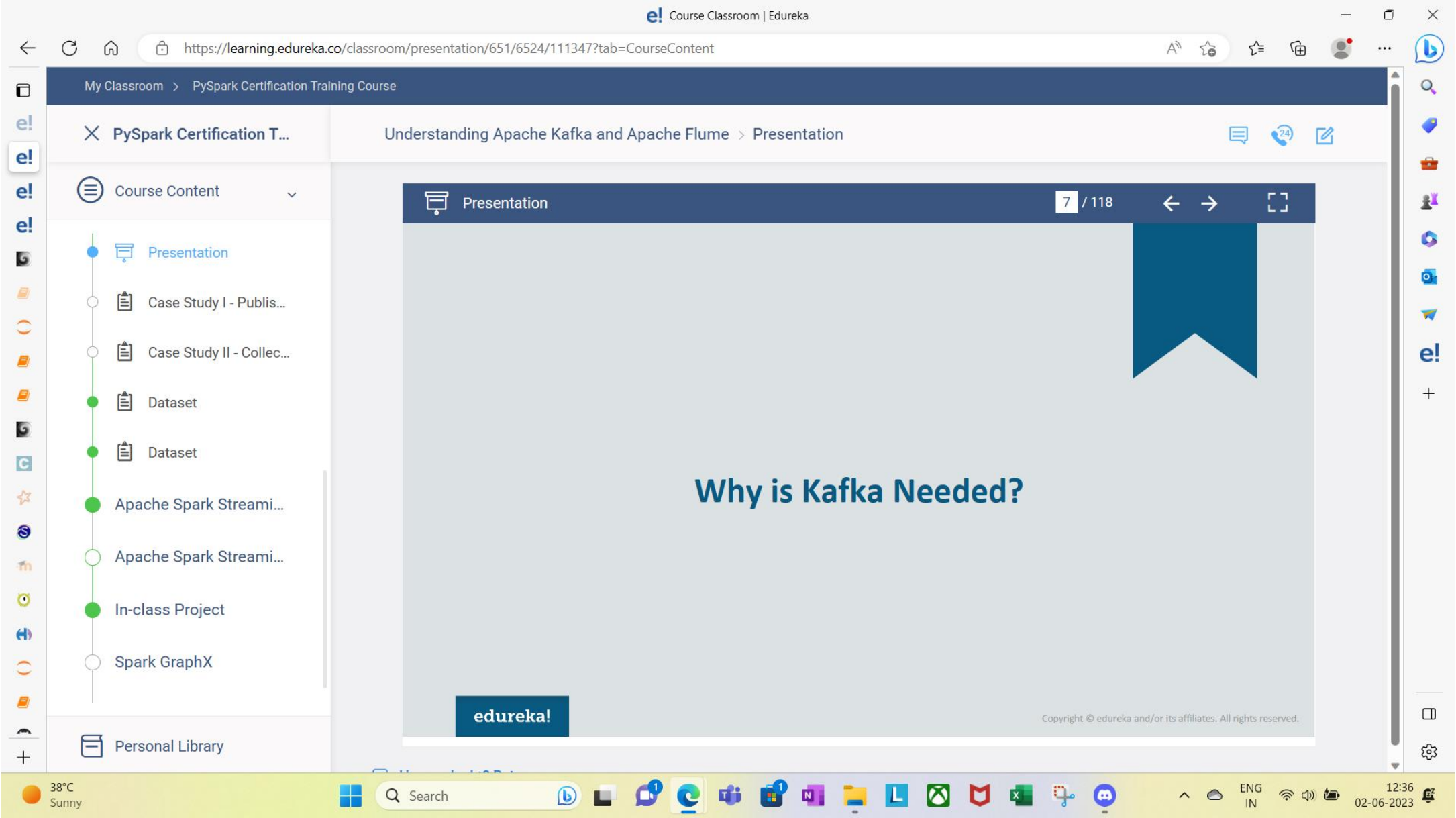
Presentation 6 / 118

Objectives

After completing this module, you should be able to:

- Understand the need for Kafka
- Describe what is Kafka
- Explain Core Concepts of Kafka
- Understand Kafka Architecture
- Understand the components of Kafka Cluster
- Configuring Kafka Cluster
- Describe what is Apache Flume?
- Describe Basic Flume Architecture
- Explain Flume Configuration and Integrate Apache Flume and Apache Kafka

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
7 / 118
Why is Kafka Needed?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

Challenges with Real-Time Data

Problem:

There are two challenges in the present real-time data:

- First is to collect the data as data is huge in amount and,
- Second is to analyse real-time data. This analysis typically includes the following types of data and much more:
  - User behaviour data
  - Application performance tracing
  - Activity data in the form of logs
  - Event messages

edureka!

Solution - Messaging Systems

Problem:

There are two challenges in the present real-time data:

- First is to collect the data as data is huge in amount and,
- Second is to analyse real-time data. This analysis typically includes the following types of data and much more:
  - User behaviour data
  - Application performance tracing
  - Activity data in the form of logs
  - Event messages

Solution:

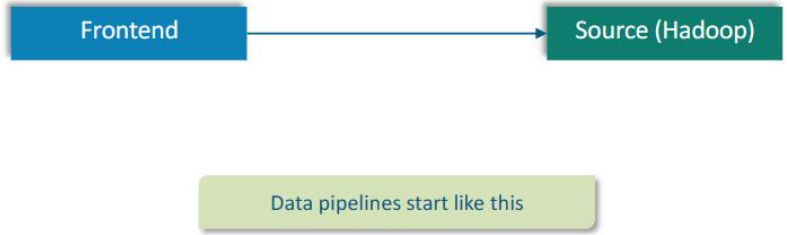
*One way to solve this problem is by **using messaging systems**, messaging systems provide seamless integration among distributed applications with the help of messages, that are shared between them.*

edureka!

Presentation 10 / 118

Why is Kafka Needed?

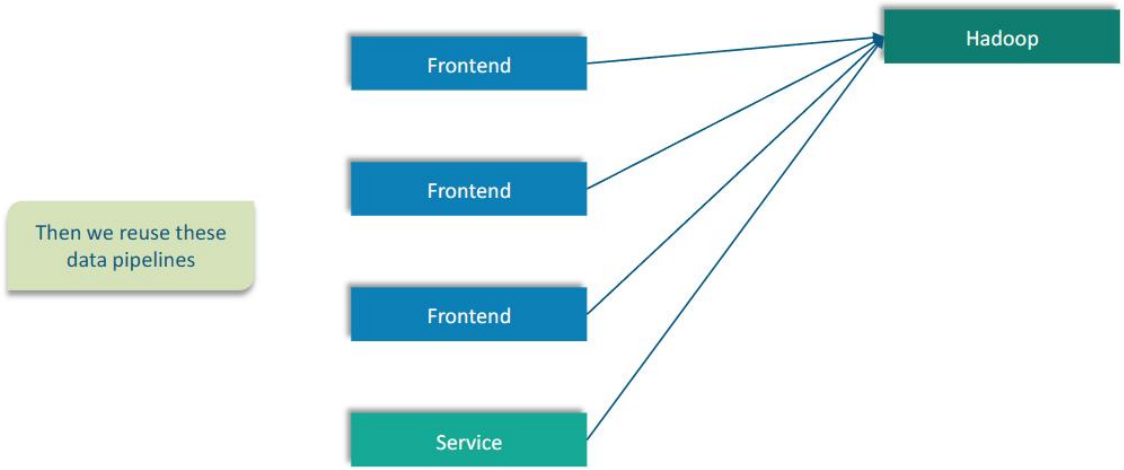
Let us see how *point to point data pipelines* makes a system more complicated.



edureka!

Presentation 11 / 118

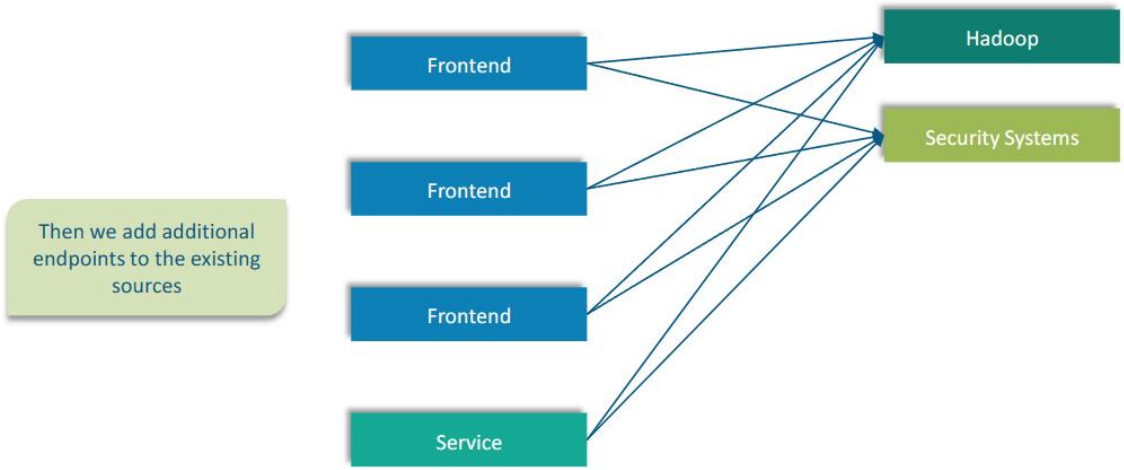
Point to Point Data Pipelines



edureka!

Presentation 12 / 118

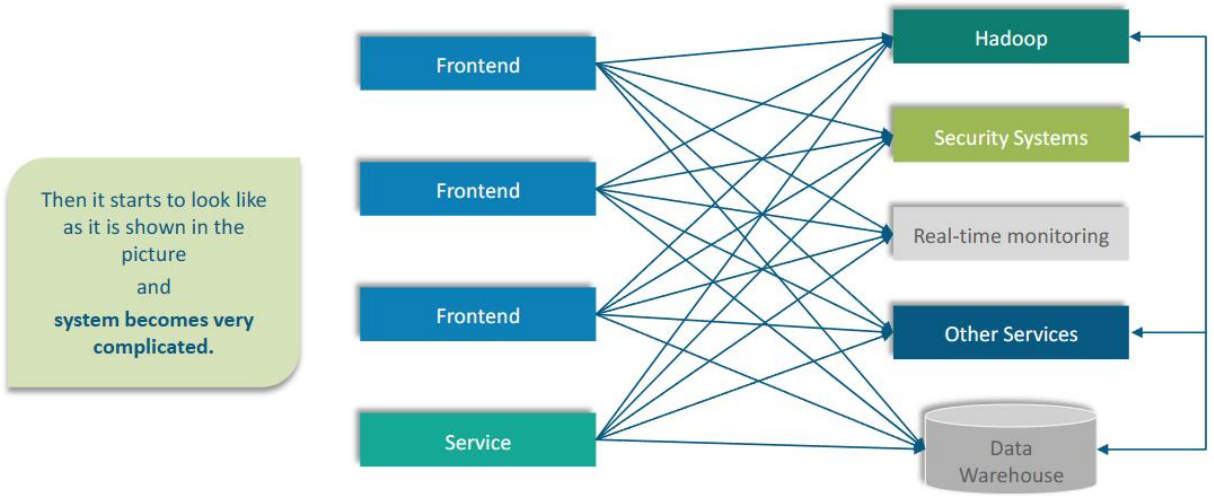
Point to Point Data Pipelines



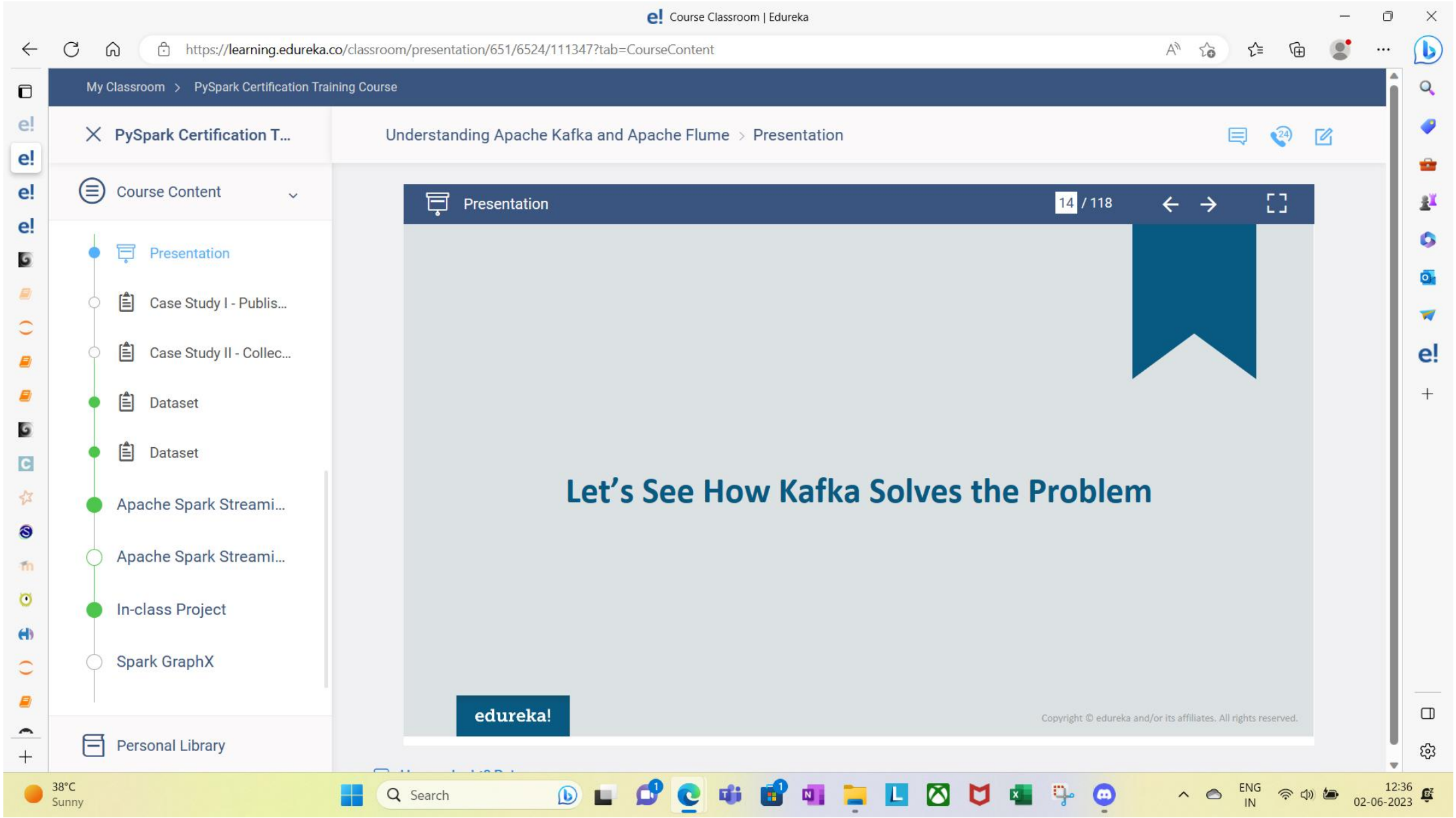
edureka!

Presentation 13 / 118

Point to Point Data Pipelines

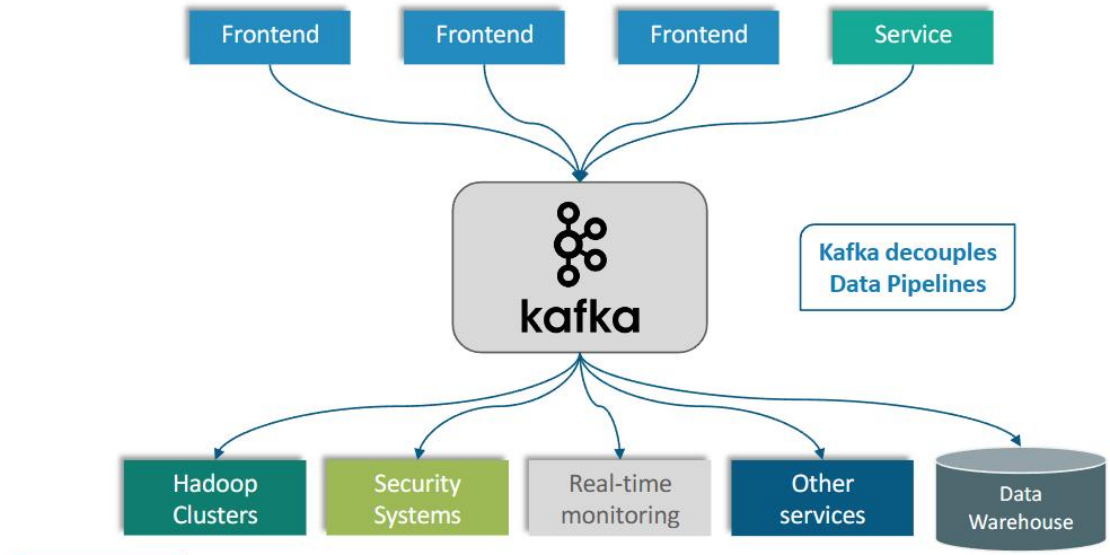


edureka!

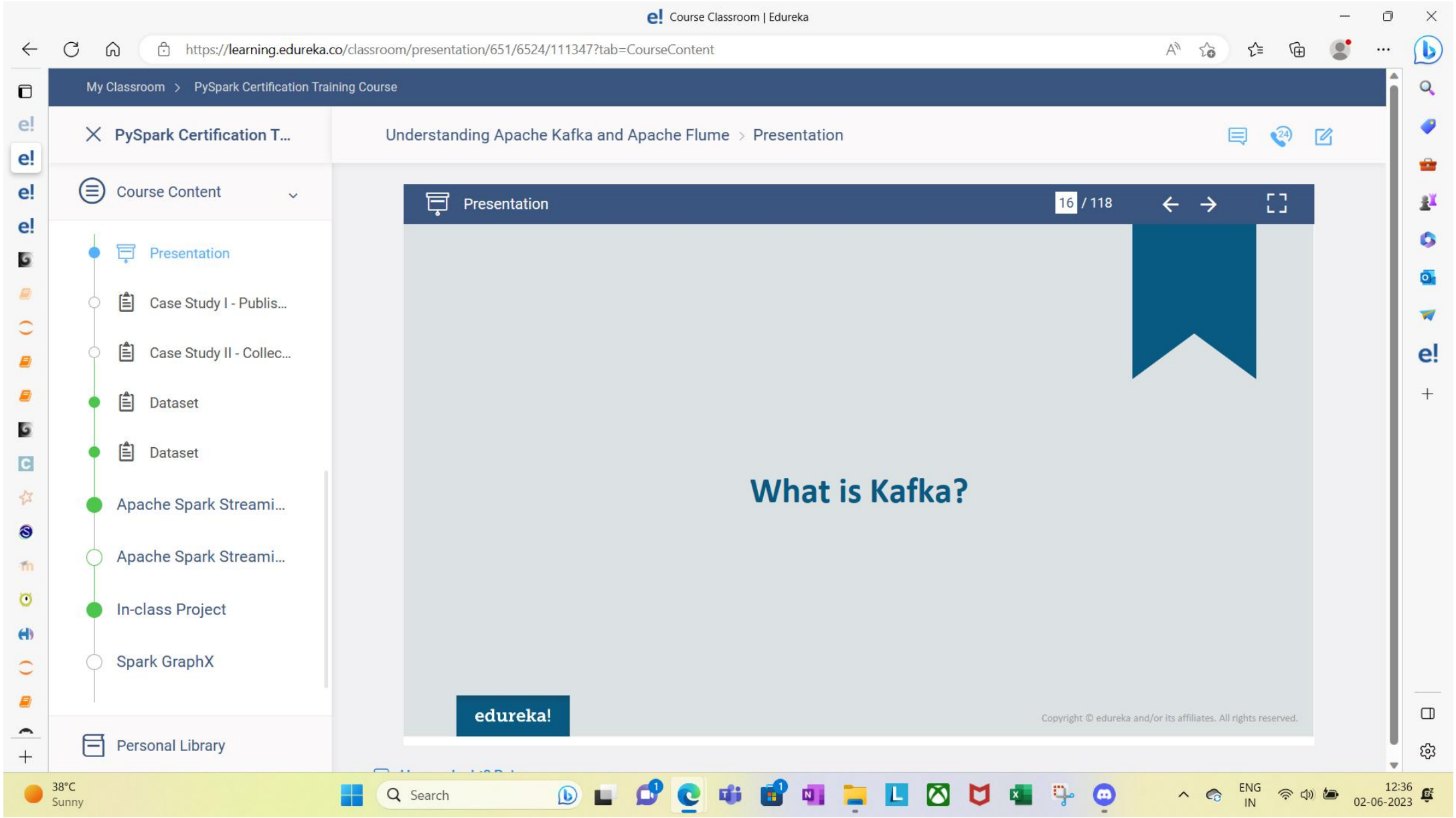

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
14 / 118
Let's See How Kafka Solves the Problem
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

Presentation 15 / 118

Kafka Decouples Data Pipelines



edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
16 / 118
What is Kafka?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

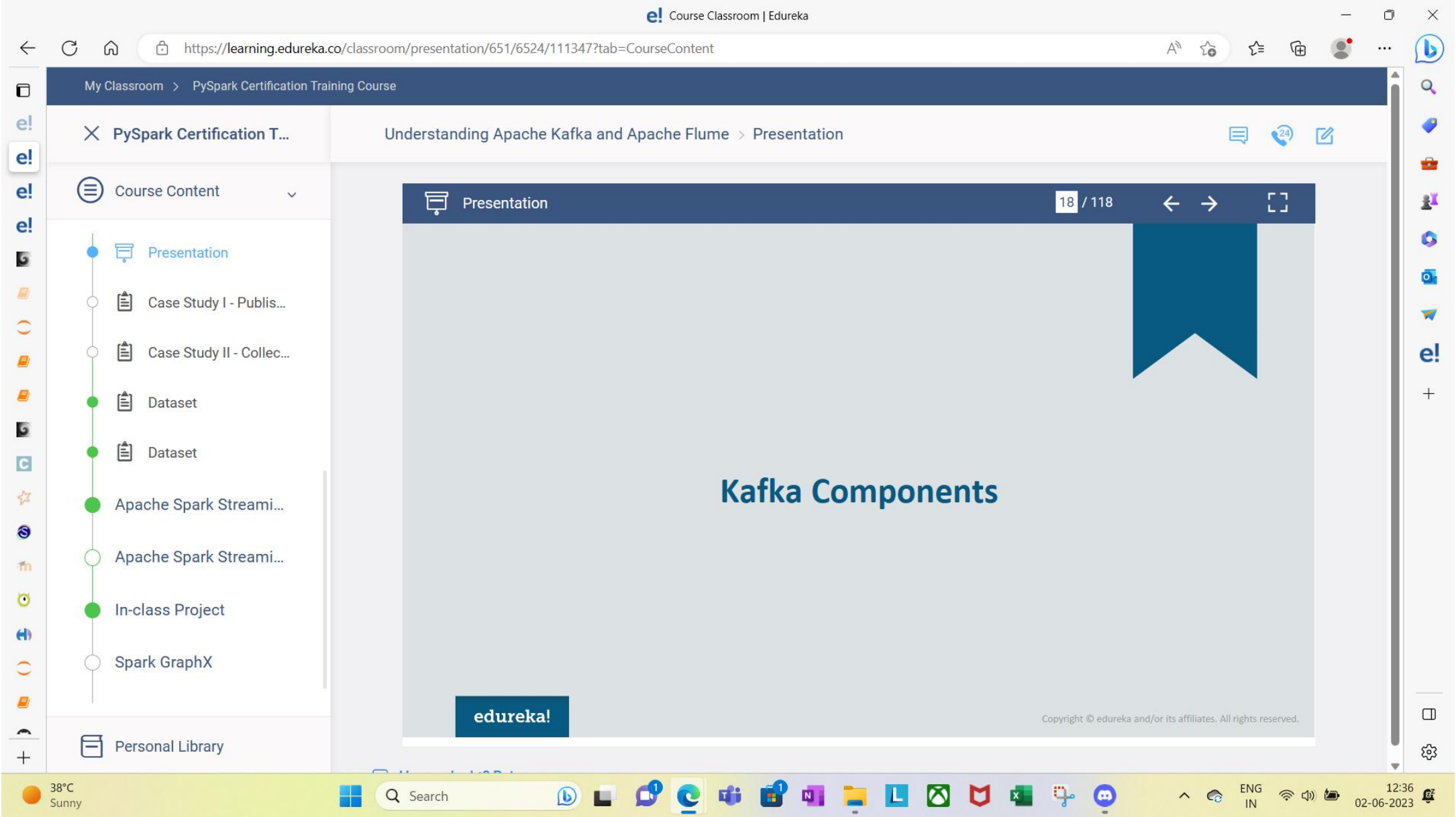
Presentation 17 / 118

What is Kafka?

- **Apache Kafka** is a distributed publish-subscribe messaging system.
- Originally developed at LinkedIn and later on became a part of Apache Project.
- Kafka is fast, scalable, durable, fault-tolerant and distributed by design.

APACHE kafka
A distributed streaming platform

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
18 / 118
Kafka Components
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

Presentation 19 / 118

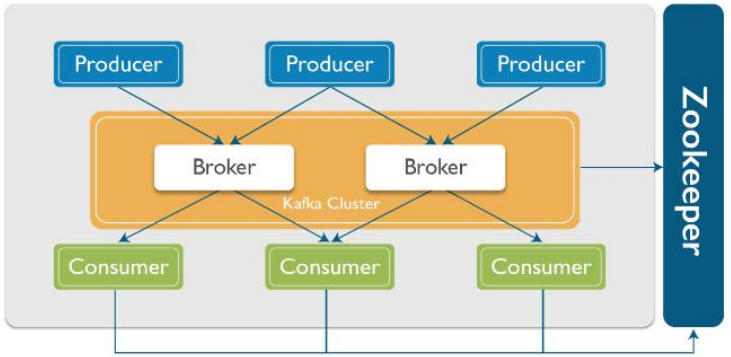
Kafka Components

Topic: A stream of messages of particular category is called a ***topic***. Producers publish messages to a topic.

Producer: A ***producer*** can be any application who can publish messages to a topic.

Consumer: A ***consumer*** can be any application that subscribe to topics and consume the messages.

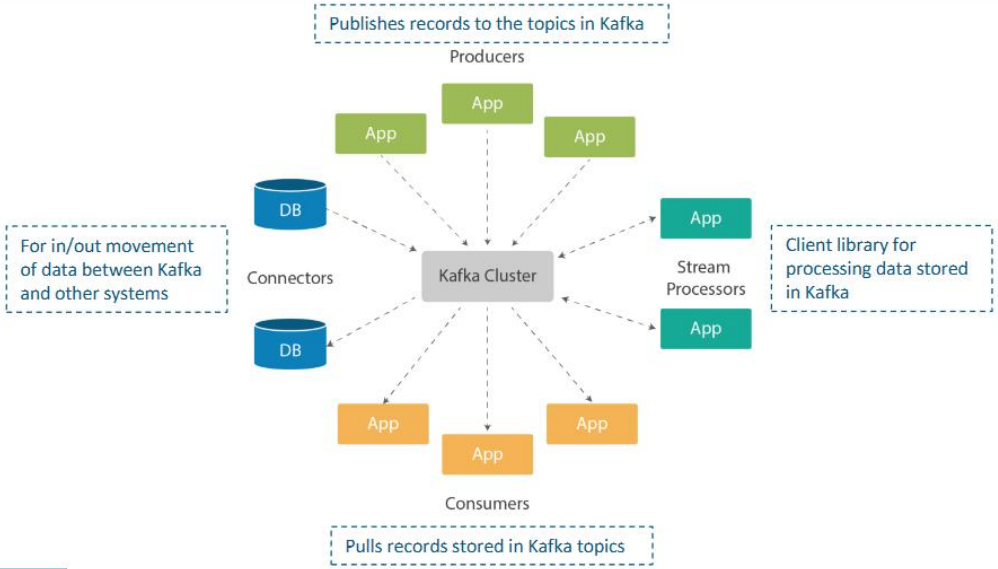
Broker: Kafka cluster is a set of servers, each of which is called a ***broker***.



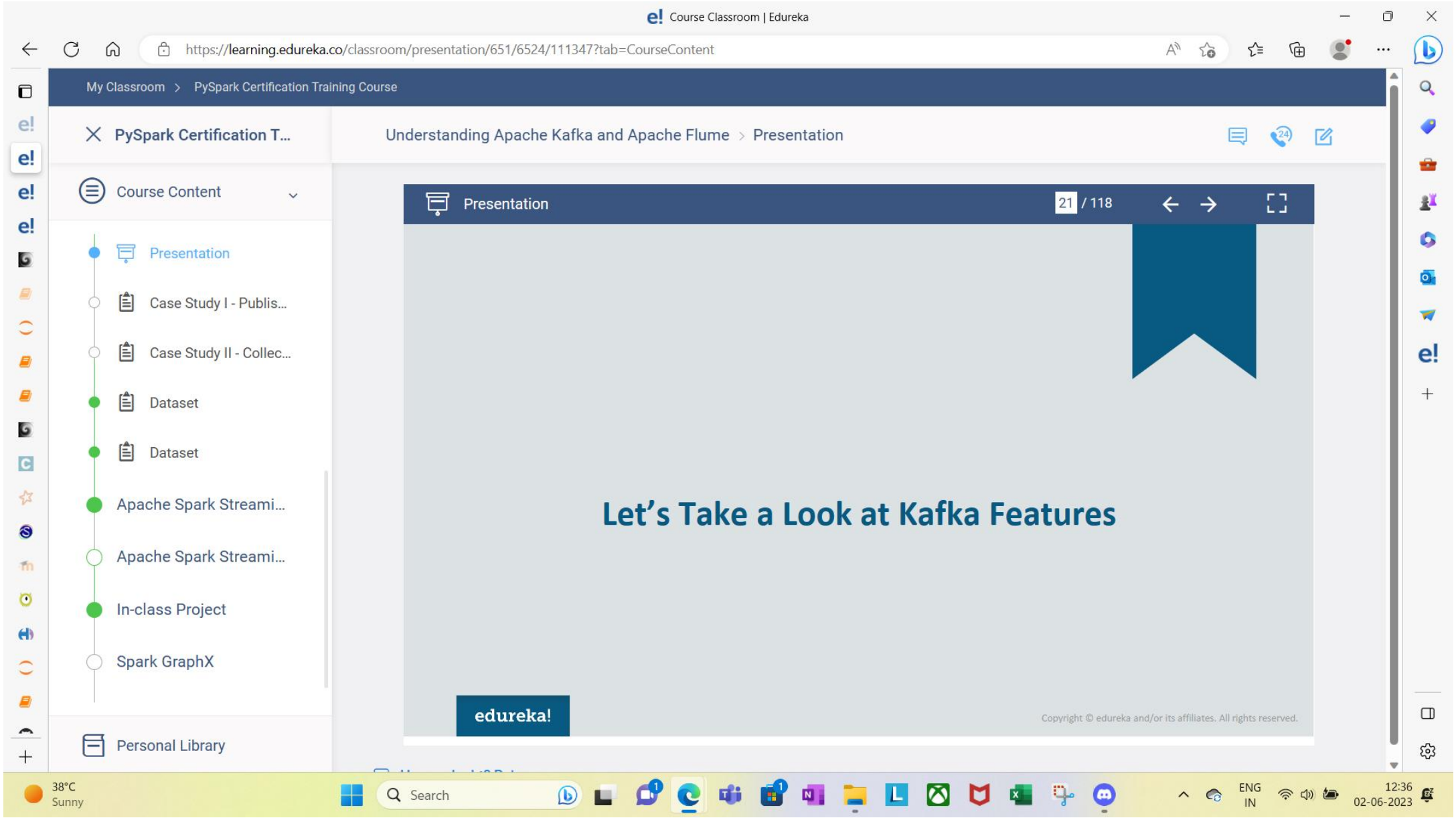
edureka!

Presentation 20 / 118

Apache Kafka Overview

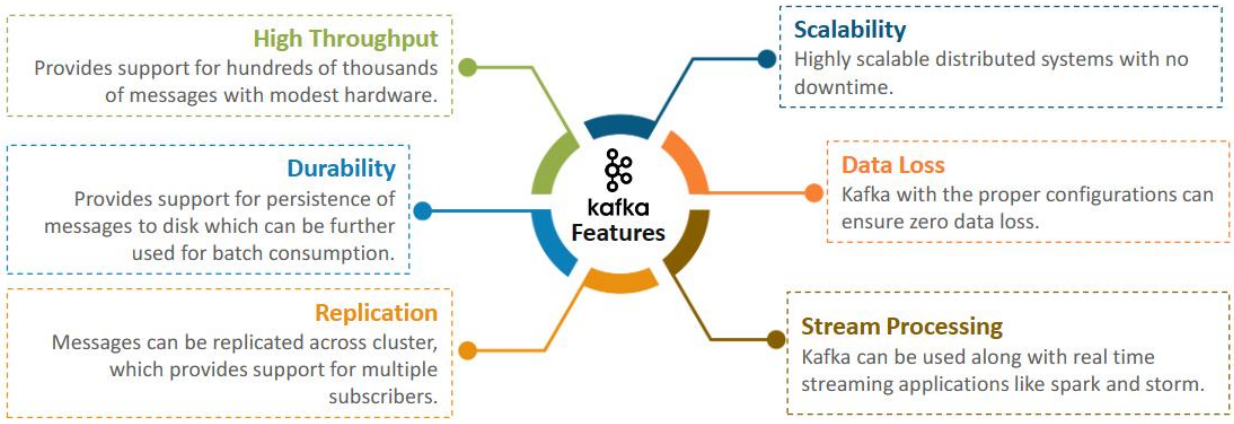


edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
21 / 118
Let's Take a Look at Kafka Features
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

Presentation 22 / 118

Kafka Features



edureka!

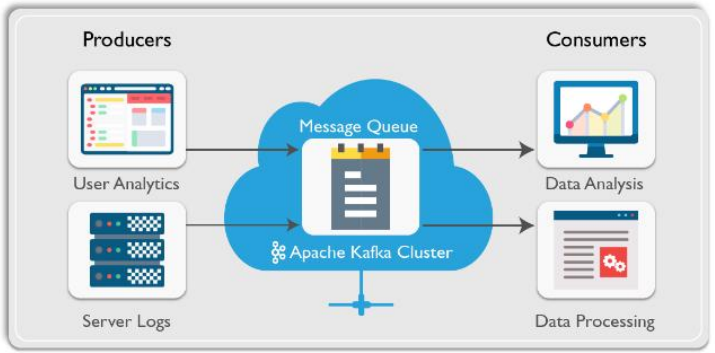
Presentation 23 / 118

Common Use Cases for Kafka

Website activity tracking: The web application sends events such as page views where they become available for real-time processing, offline analytics in Hadoop.

Operational metrics: Alerting and reporting on operational metrics.

Log aggregation: Kafka can be used to collect logs from multiple services.



Stream processing: A framework such as Spark Streaming reads data from a topic, processes it and writes processed data to a new topic where it becomes available for users and applications.

edureka!

Presentation 24 / 118

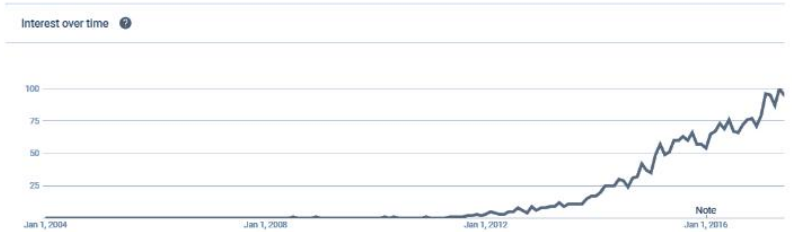
Kafka Growth Exploding

- More than **1/3** of all Fortune **500** companies use **Kafka**.
- These companies includes the top ten travel companies, **7** of top ten banks, **8** of top ten insurance companies, **9** of top ten telecom companies.
- **LinkedIn**, **Microsoft** and **Netflix** process billions of messages a day with Kafka (1,000,000,000,000).
- **Kafka** is used for **real-time streams** of data & used to collect big data for **real time analysis**.

86% of respondents reported that the number of their systems that use Kafka is increasing

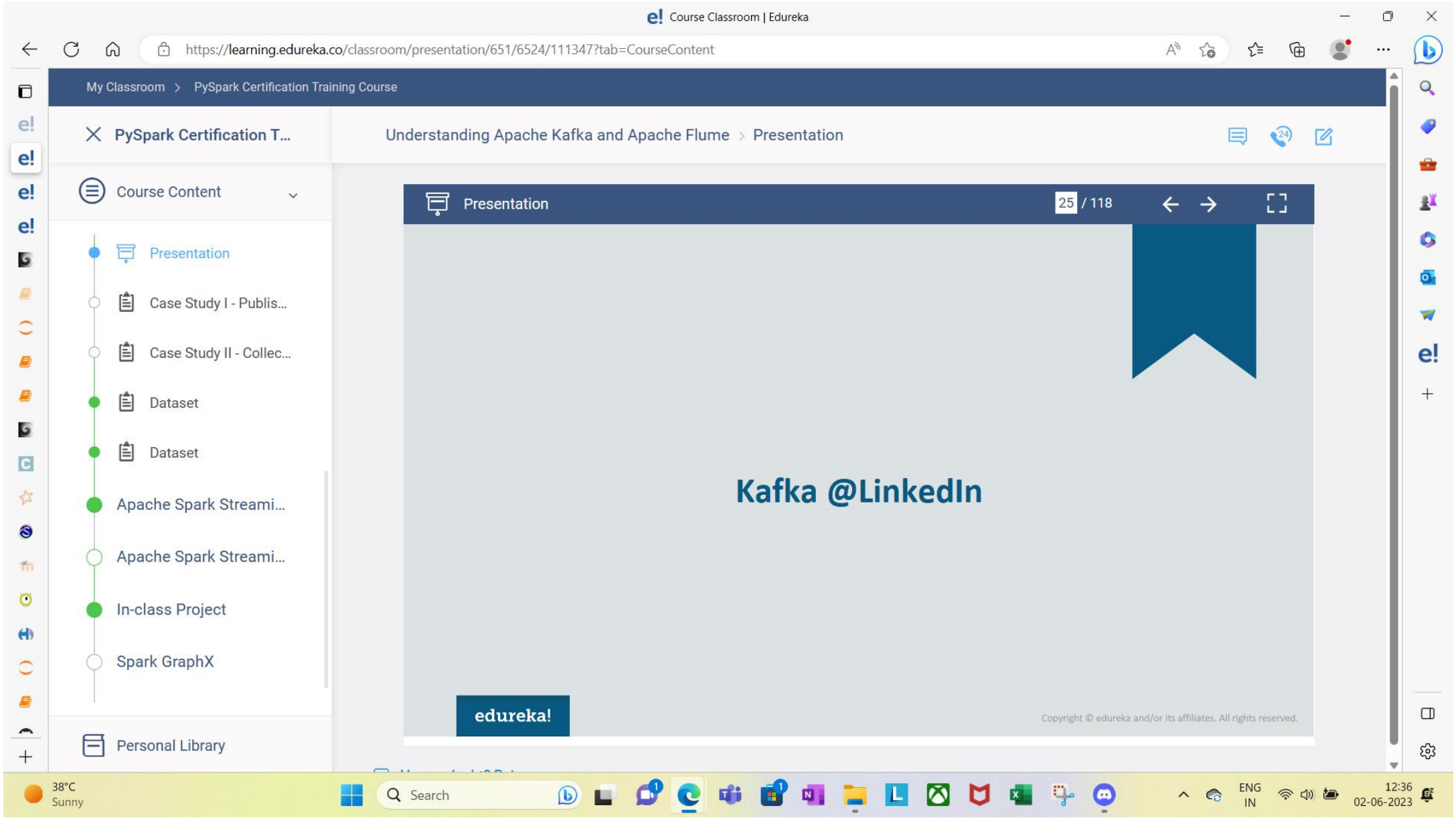
20% reported that the number is "growing a lot!"

52% of organizations have at least 6 systems running Kafka.



Source: Google Trends

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
25 / 118
Kafka @LinkedIn
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

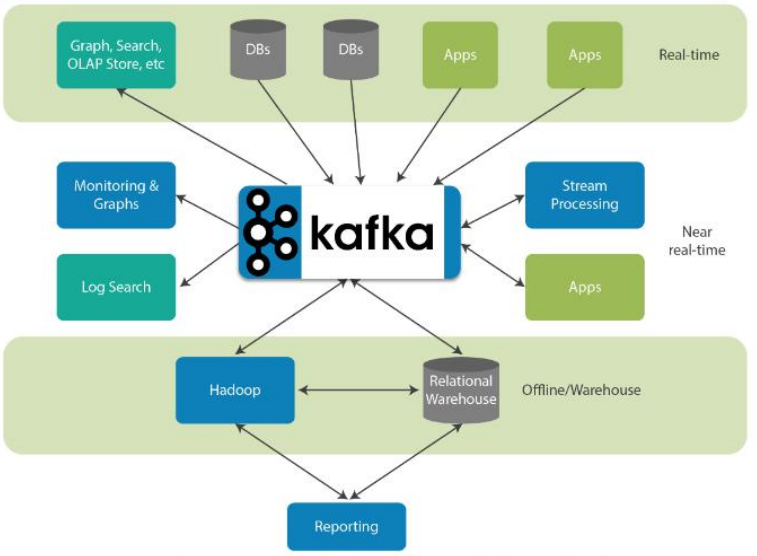
Presentation 26 / 118

Kafka @LinkedIn

- *1100+ commodity machines*
- *31,000+ topics*
- *350,000+ partitions*

- *675 billion messages/day*
- *150 TB/day in*
- *580 TB/day out*

Peak Load
- *10.5 million messages/sec*
- *18.5 GB/sec Inbound*
- *70.5 GB/sec Outbound*

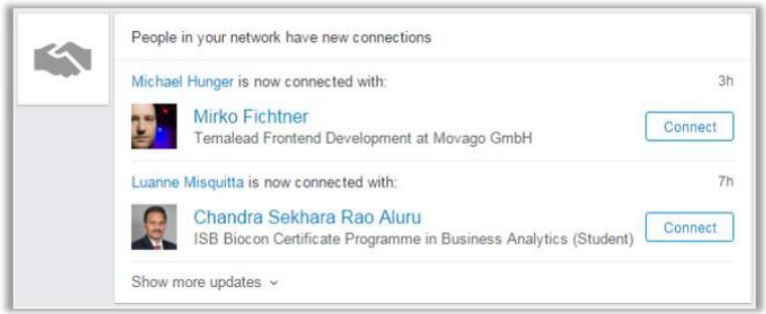


A modern stream-centric data architecture built around Kafka

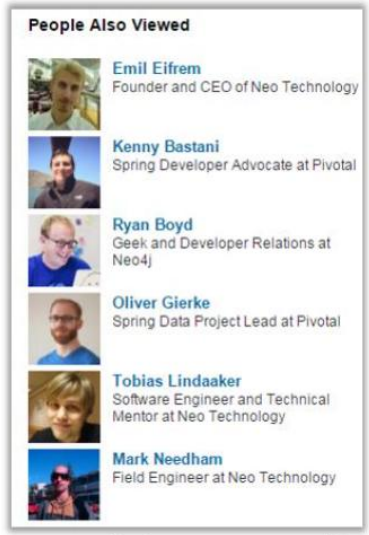
edureka!

Presentation 27 / 118

Kafka @LinkedIn



LinkedIn Newsfeed is powered by Kafka



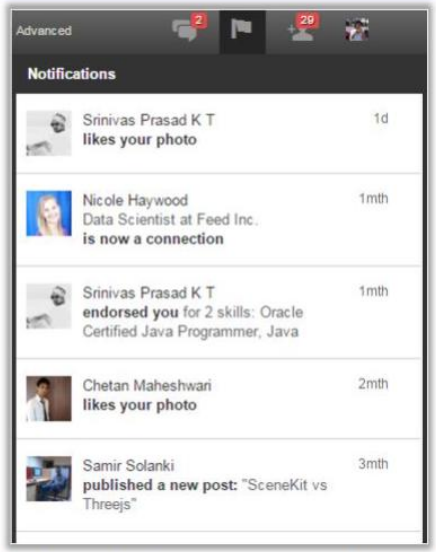
LinkedIn recommendations are powered by Kafka

edureka!

Presentation 28 / 118

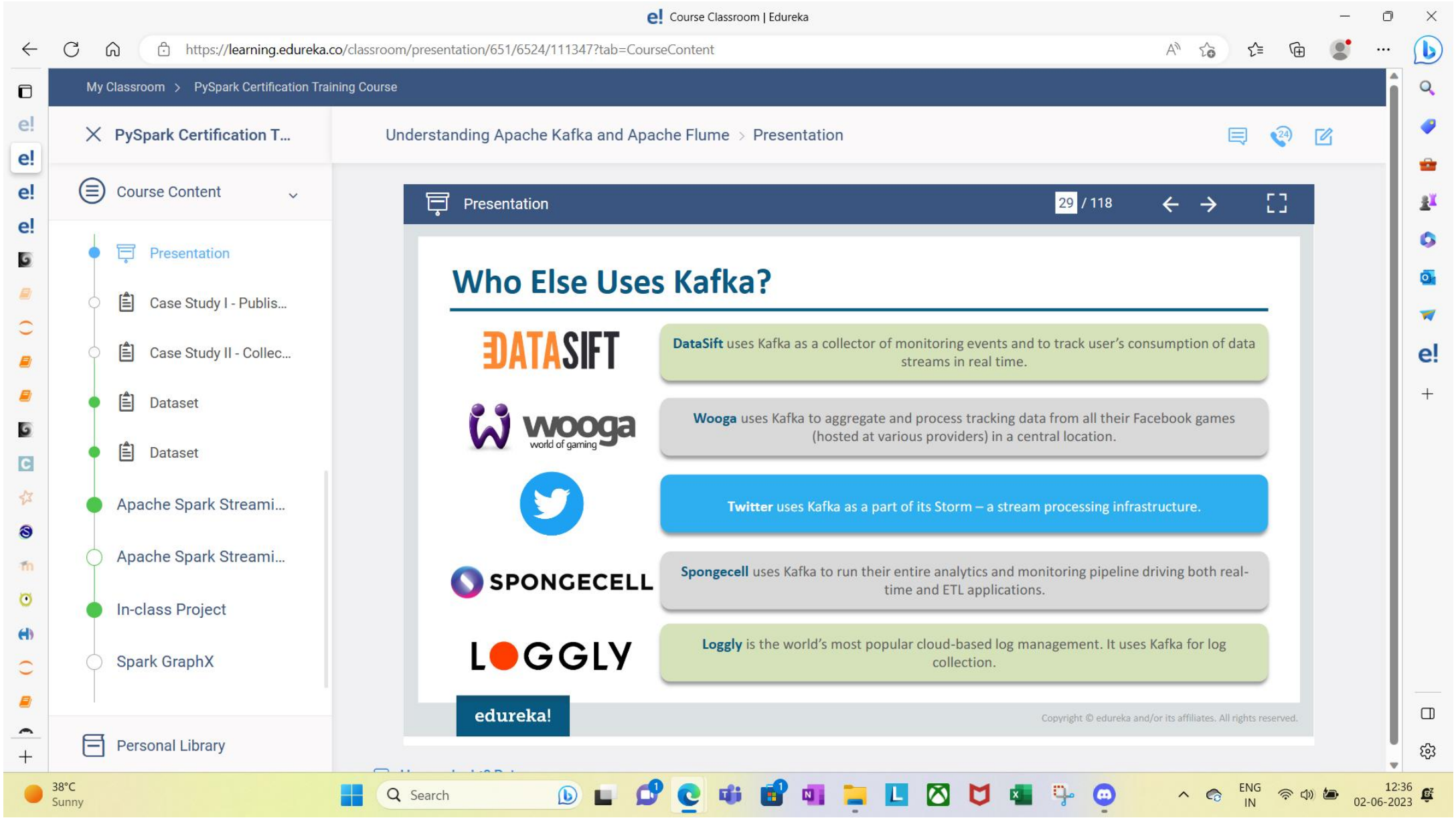
Kafka @LinkedIn

Apart from this LinkedIn uses Kafka for many other purposes like log monitoring, performance metrics, search improvement etc.



LinkedIn notifications are powered by Kafka

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
29 / 118
Who Else Uses Kafka?
DATASIFT
DataSift uses Kafka as a collector of monitoring events and to track user's consumption of data streams in real time.
wooga
world of gaming
Wooga uses Kafka to aggregate and process tracking data from all their Facebook games (hosted at various providers) in a central location.
Twitter uses Kafka as a part of its Storm – a stream processing infrastructure.
SPONGECELL
Spongecell uses Kafka to run their entire analytics and monitoring pipeline driving both real-time and ETL applications.
LOGGLY
Loggly is the world's most popular cloud-based log management. It uses Kafka for log collection.
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

Presentation 30 / 118

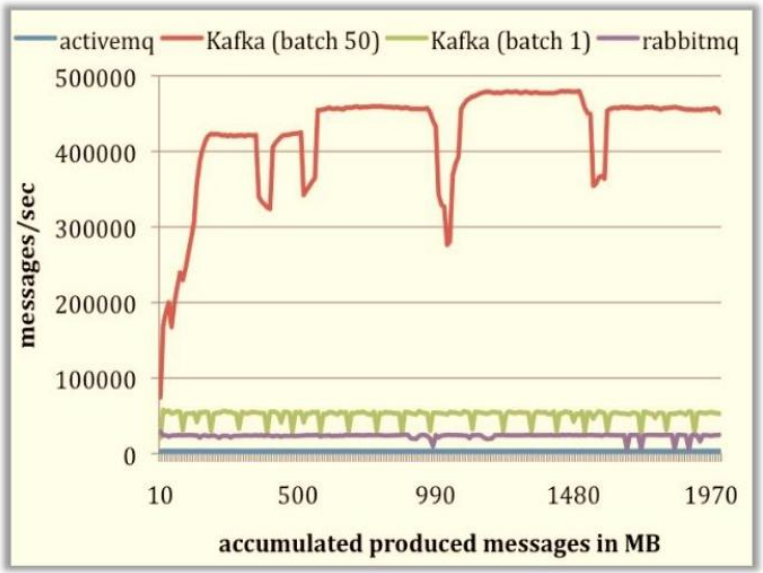
edureka!

Presentation 31 / 118

Comparing Messaging Systems

Kafka has a more efficient storage format.

On average, each message had an overhead of 9 bytes in Kafka, versus 144 bytes in ActiveMQ.

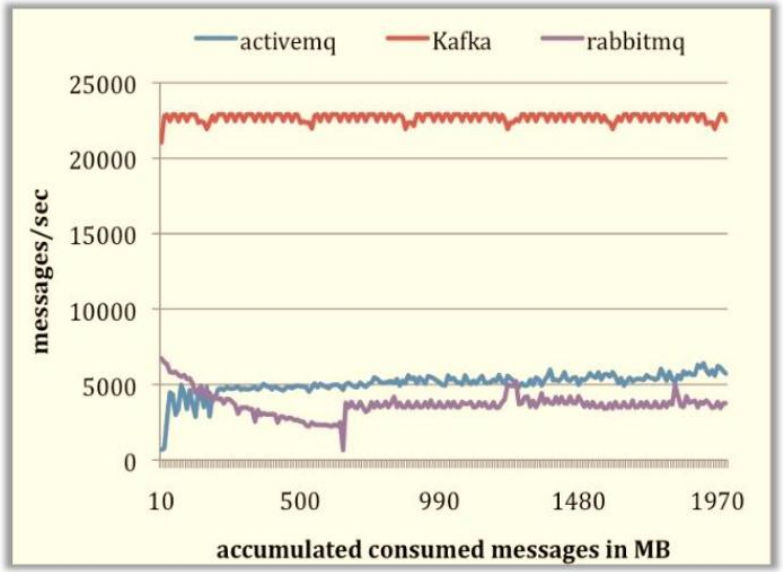


edureka!

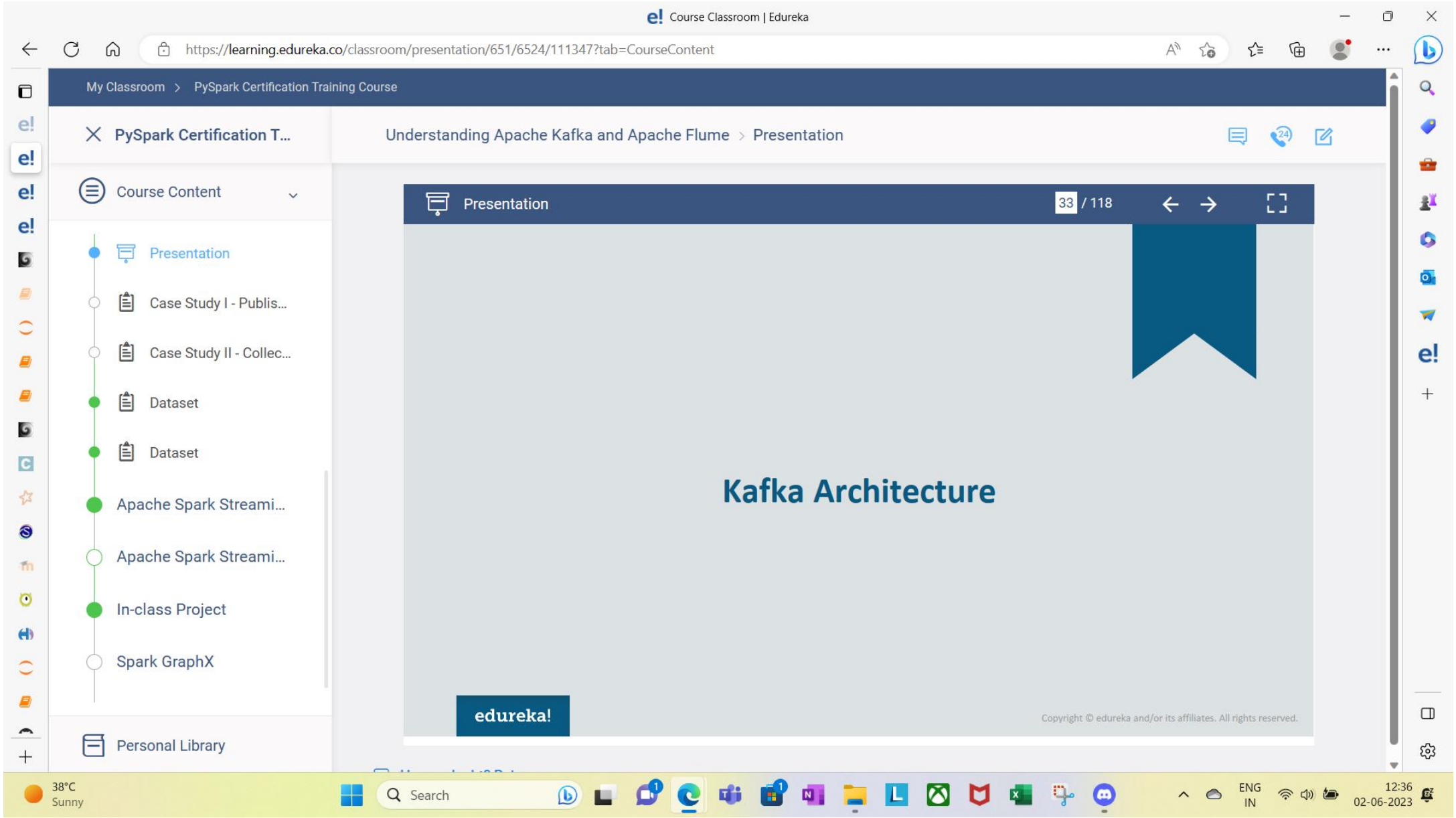
Presentation 32 / 118

Comparing Messaging Systems

In both ActiveMQ and RabbitMQ brokers maintain delivery state of every message by writing to disk but in case of Kafka there is no disk write which makes it fast.

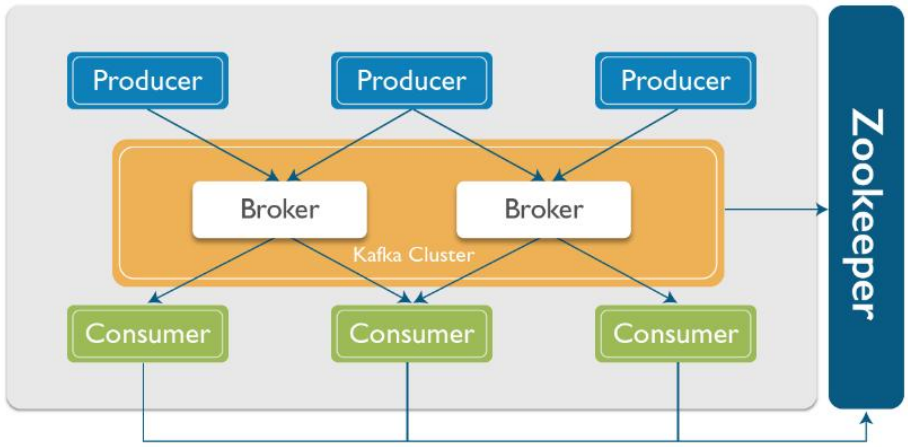


edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
33 / 118
Kafka Architecture
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:36
02-06-2023

Presentation 34 / 118

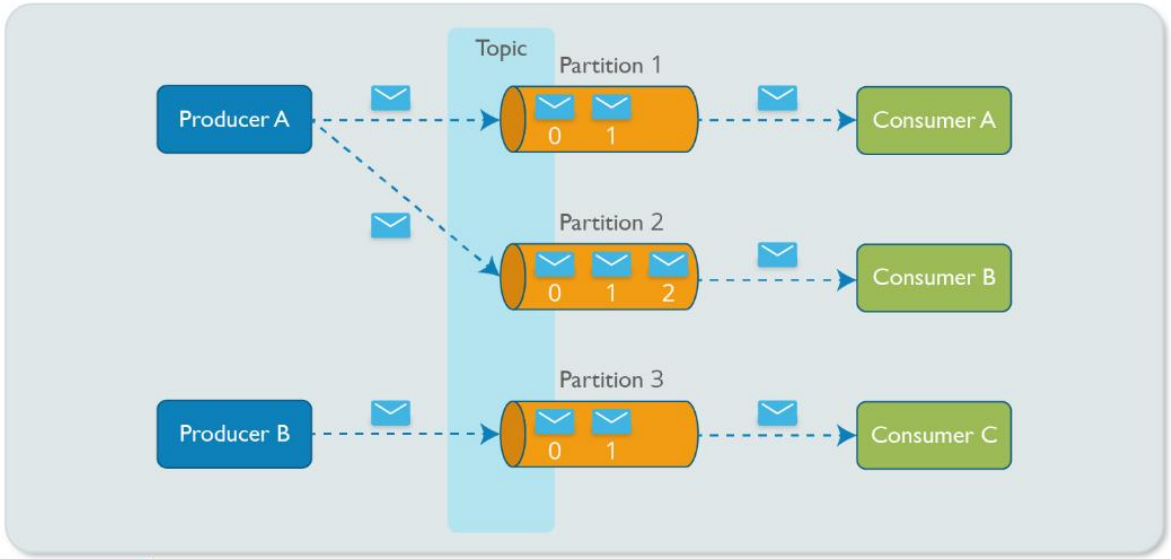
Recap : Kafka Architecture



edureka!

Presentation 35 / 118

Kafka Architecture - Deep Dive

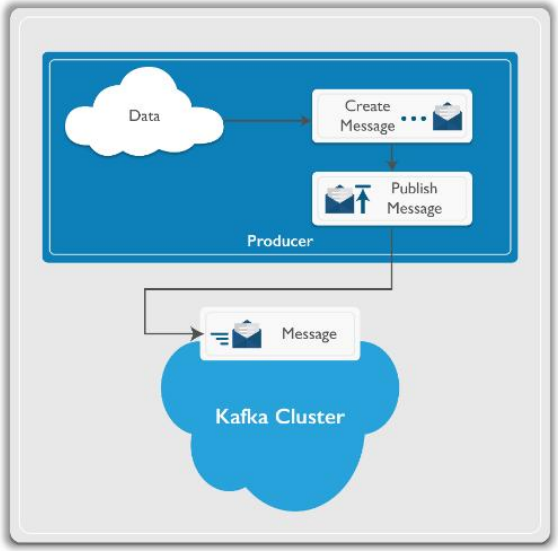


edureka!

Presentation 36 / 118

Kafka Producer

- *Applications publishes messages* to the *topic in Kafka cluster*.
- While *writing* messages, it is also possible to *attach a key with* the *message*.
- By attaching key the *producers* basically *provide* a *guaranty* that all *messages* with the *same key* will *arrive* in the *same partition*.
- Supports both *async* (less durable) and *sync* modes.
- Publishes as much messages as fast as the broker in a cluster can handle.

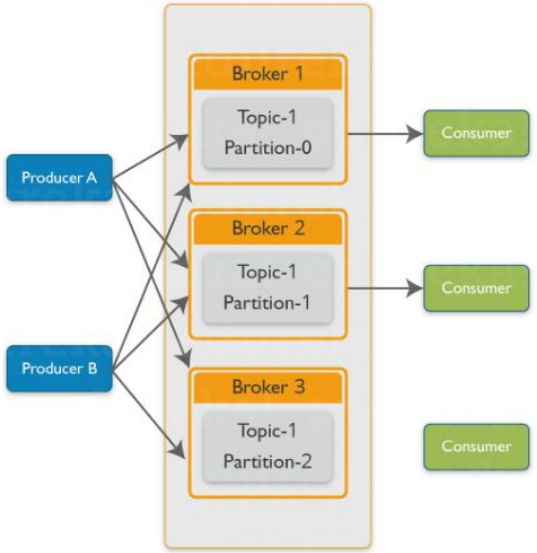


edureka!

Presentation 37 / 118

Kafka Broker

- Kafka cluster basically comprised of one or more servers.
- *Each* of the *servers* in the *cluster* is called a **broker**.
- *Handles hundreds* of *megabytes* of *writes* from *producers* and *reads* from *consumers*.
- *Retains* all *published messages irrespective* of whether it is *consumed* or *not*.
- If retention is configured for n days ,then messages once published it is available for consumptions for configured n days and thereafter it is discarded.
- *Works like* a *queue* if *consumer* instances *belong* to *same consumer group*, else works like *publish-subscribe*.

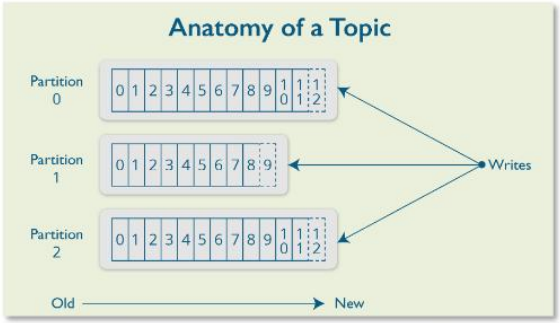


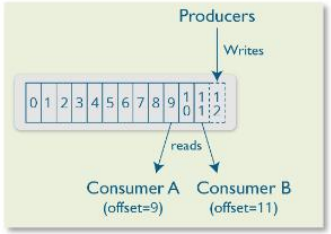
edureka!

Presentation 38 / 118

Topics and Partitions

- A *topic* is a *category* or *feed name* to which *records are published*.
- *Topics* are *broken* up into *ordered commit logs* called *partitions*.
- Each *message* in a *partition* is assigned a *sequential id* called an *offset*.
- Data is retained for a *configurable period of time*.
- *Writes* to a partition *are* generally *sequential* thereby *reducing* the *number of hard disk seeks*.
- *Reading messages* from partition can either be from *beginning* and *also* can *rewind* or *skip* to any *point* in partition by *supplying* an *offset value*.



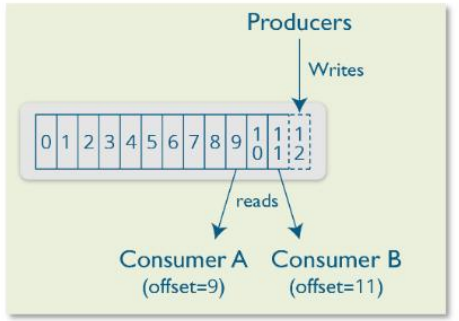


edureka!

Presentation 39 / 118

Message Ordering & Guarantees

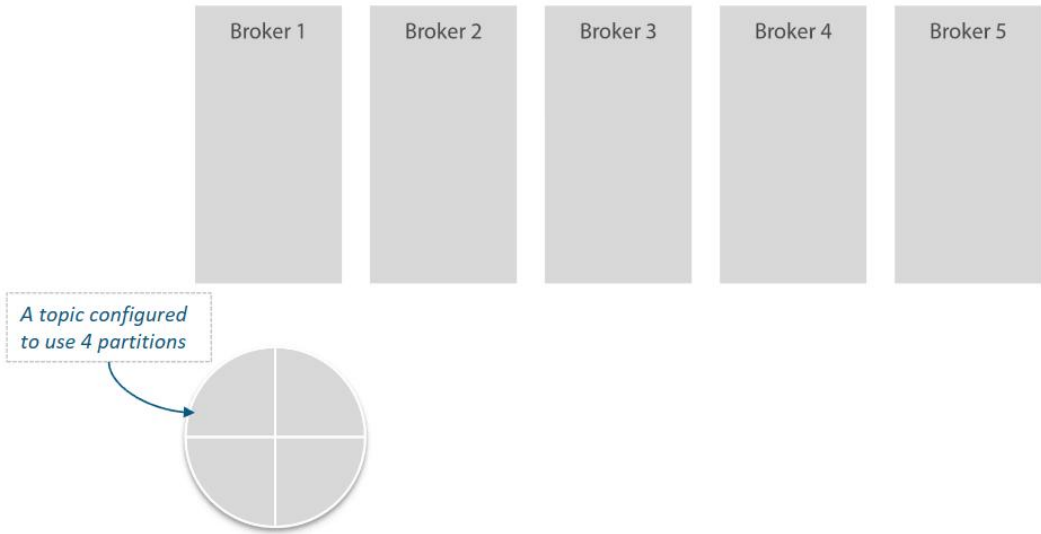
- *Ordering* is only *guaranteed* within a *partition for a topic*.
- To ensure ordering:
  - *Group messages in a partition by key (producer).*
  - *Configure exactly one consumer instance per partition within a consumer group.*
- *Messages sent* by a *producer* to a particular topic *partition* will be *appended in the order* they are sent.
- A *consumer instance* sees messages in the *order they are stored in* the *log*.
- For a *topic with replication factor N,* Kafka can *tolerate* up to *N-1 server failures* without "losing" any messages committed to the log.



edureka!

Presentation 40 / 118

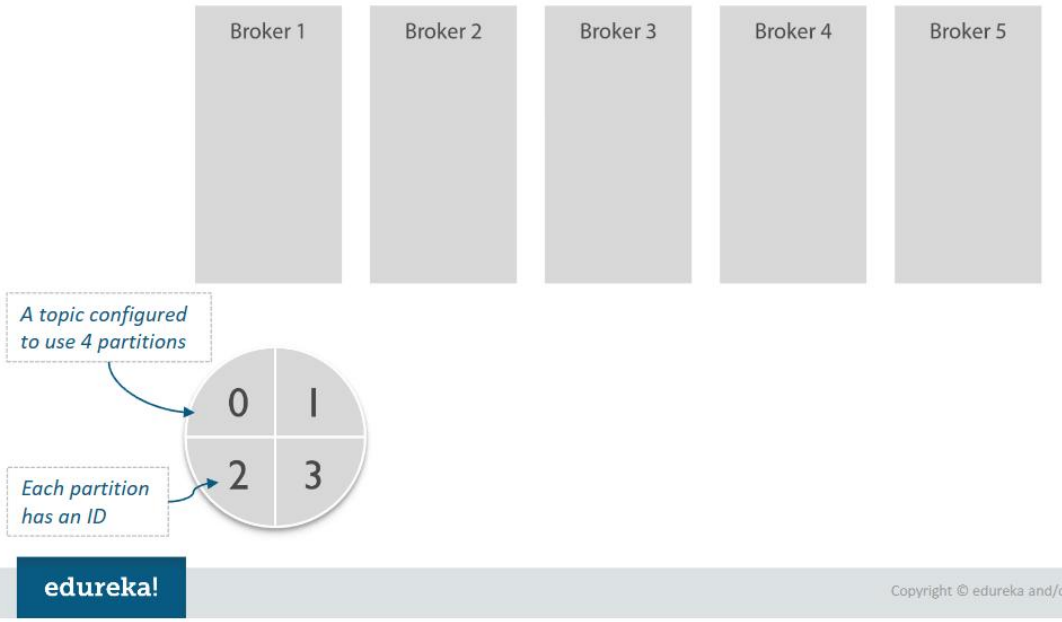
Topics vs. Partitions vs. Replicas



edureka!

Presentation 41 / 118

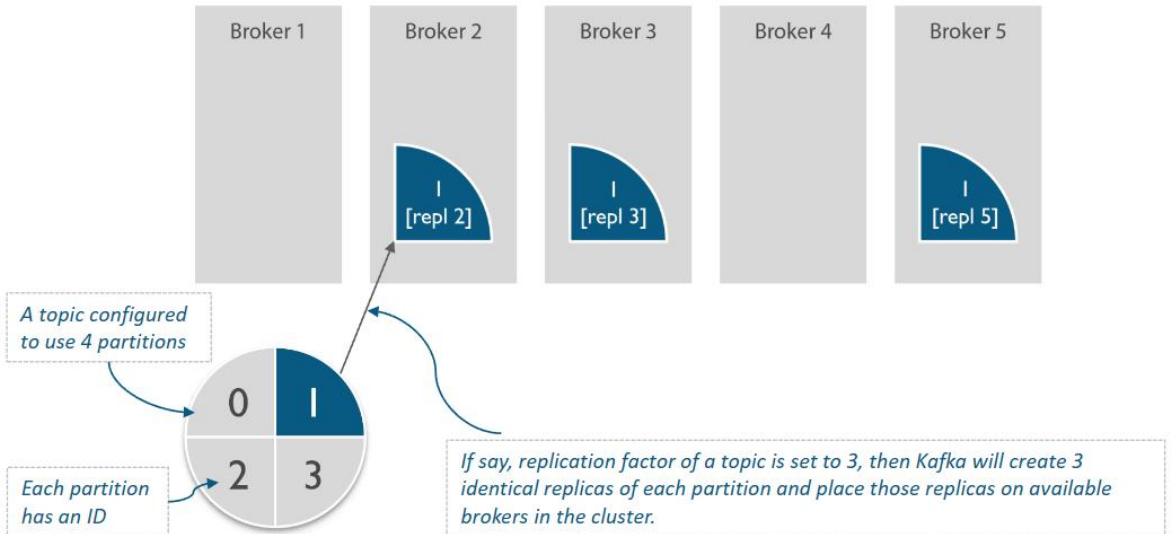
Topics vs. Partitions vs. Replicas



edureka!

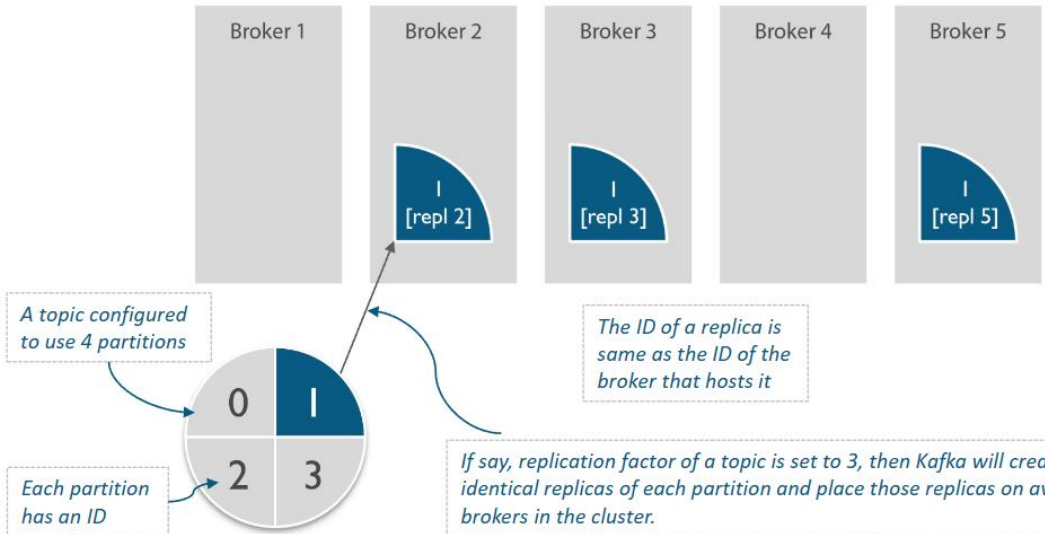
Presentation 42 / 118

Topics vs. Partitions vs. Replicas



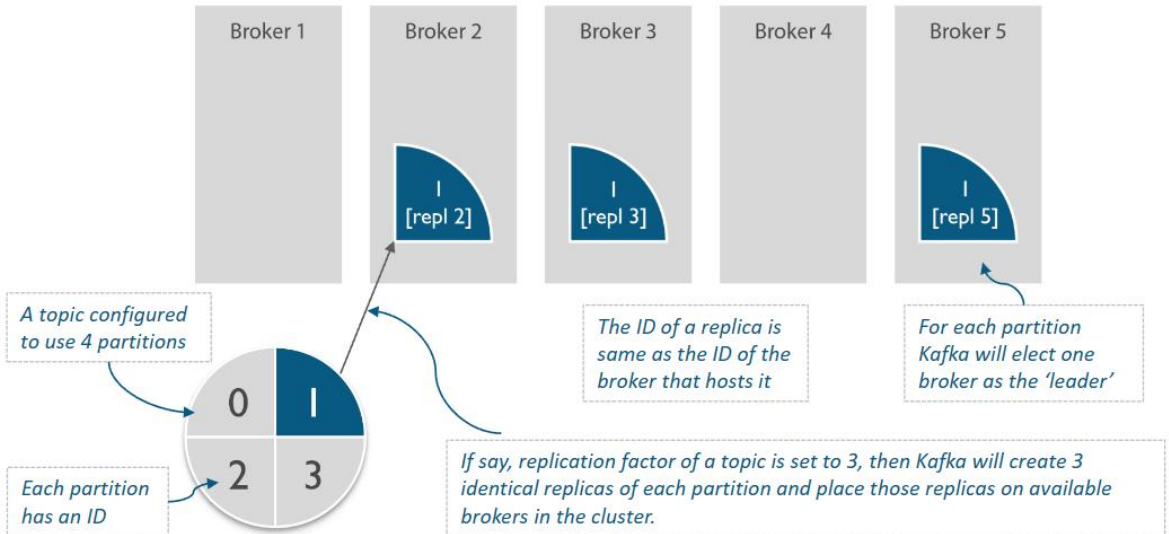
edureka!

Topics vs. Partitions vs. Replicas



edureka!

Topics vs. Partitions vs. Replicas



edureka!

Presentation 45 / 118

Kafka Broker Configurations

The essential configurations are the following:

- broker.id
- log.dirs.
- zookeeper.connect

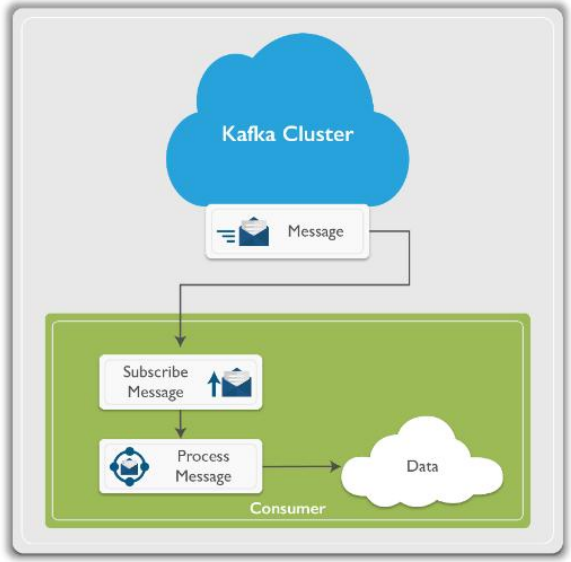
Name	Description	Default Value
broker.id	Each broker is uniquely identified by a non-negative integer ID.	0
zookeeper.connect	This specifies the ZooKeeper's connection string in the hostname:port/chroot form. Here, chroot is a base directory that is prepended to all path operations.	localhost:2181
log.dirs	These are the directories in which the log data is stored.	/tmp/kafka-logs
default.replication.factor	default replication factors for automatically created topics	1
num.partitions	The default number of log partitions per topic	1
host.name	Only used when `listeners` is not set. This is the hostname of broker.	NULL

edureka!

Presentation 46 / 118

Kafka Consumer

- *Applications subscribes* and *consumes messages* from *brokers* in *Kafka cluster*.
- During *consumption of messages* from a *topic* a *consumer group* can be configured with *multiple consumers*.
- Each *consumer* of *consumer group* reads *messages* from a *unique subset of partitions* in each topic they subscribe to.
- *Messages* with *same key* arrives at *same consumer*.
- Supports both Queuing and Publish-Subscribe.
- *Consumers* have to *maintain* the *number of messages consumed*.

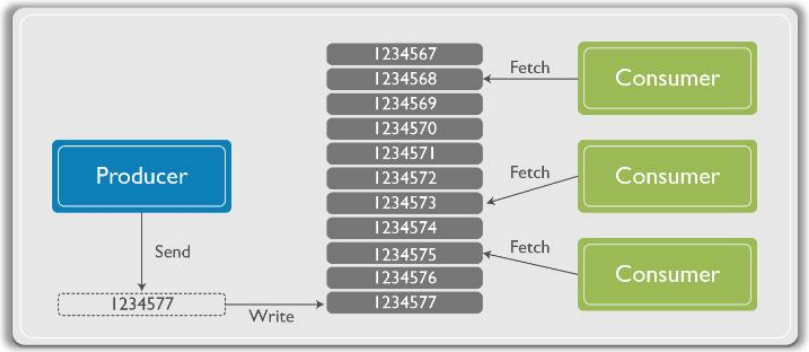


edureka!

Presentation 47 / 118

Kafka Consumer

- *Multiple consumers* can *read* from the *same topic*.
- *Each consumer* is *responsible* for *managing it's own offset*.
- *Messages stay* on *Kafka* i.e. they are not removed after they are consumed.



edureka!

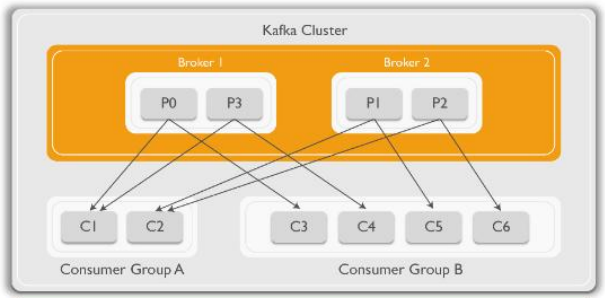
Presentation 48 / 118

Consumer - Groups

Consumers can be *organised* into *consumer groups*.

Common Patterns:

- *All consumer instances in one group*
  - Acts like a *traditional queue* with load balancing.
- *All consumer instances in different groups*
  - *All messages* are *broadcast* to *all consumer instances*.
- *"Logical-Subscriber" – Many consumer instances in a group*
  - Consumers are *added* for *scalability* and *fault tolerance*.
  - *Each consumer instance* reads from one or more partitions for a topic.
  - There *cannot* be more *consumer instances* than *partitions*.

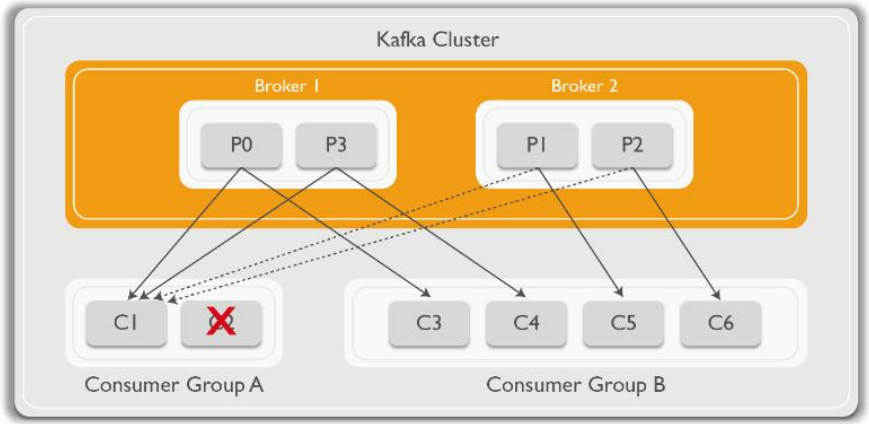


Consumer groups provide isolation to topics and partition.

edureka!

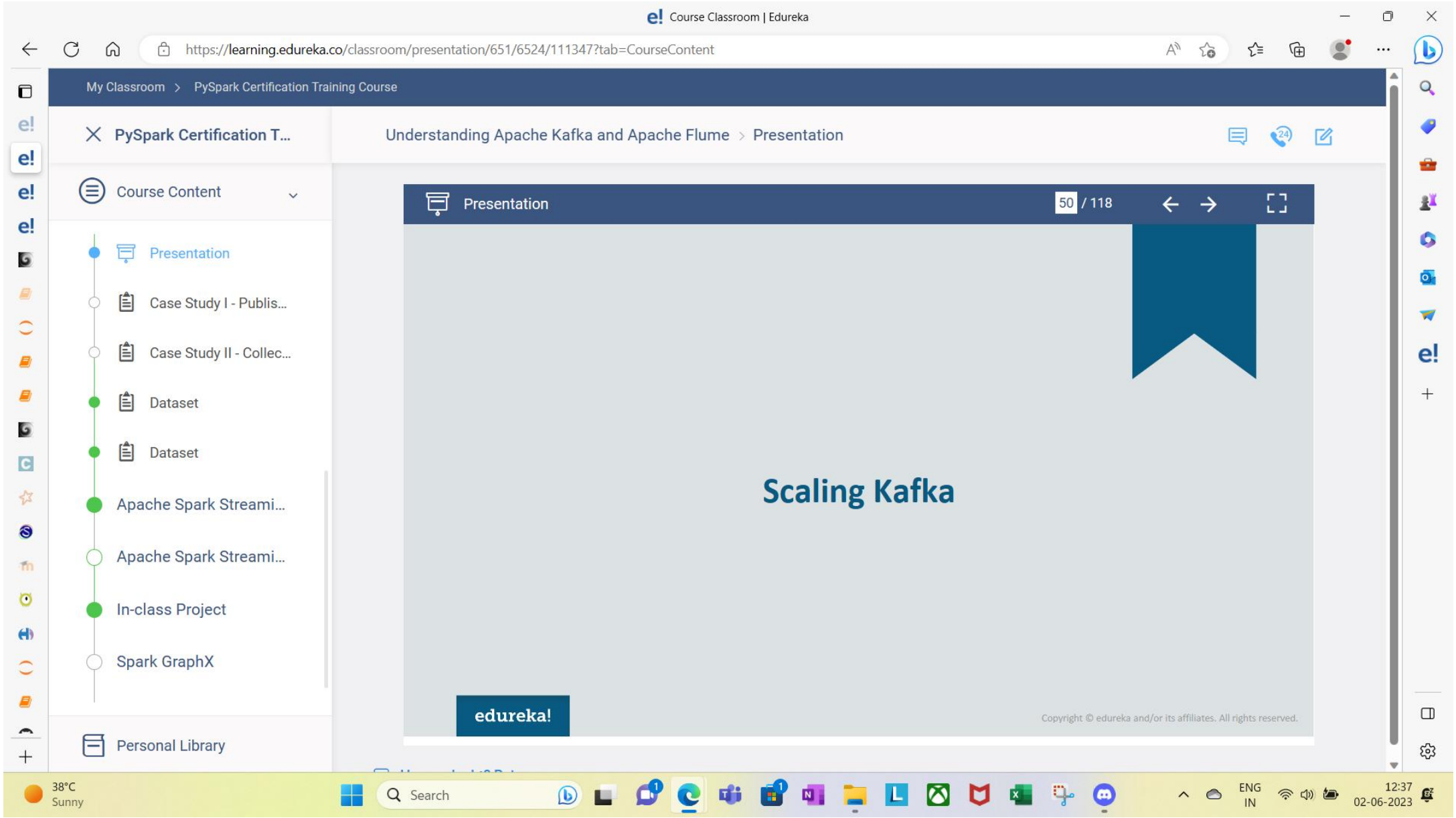
Presentation 49 / 118

Consumer - Groups



Consumers can rebalance themselves for partitions

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
50 / 118
Scaling Kafka
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

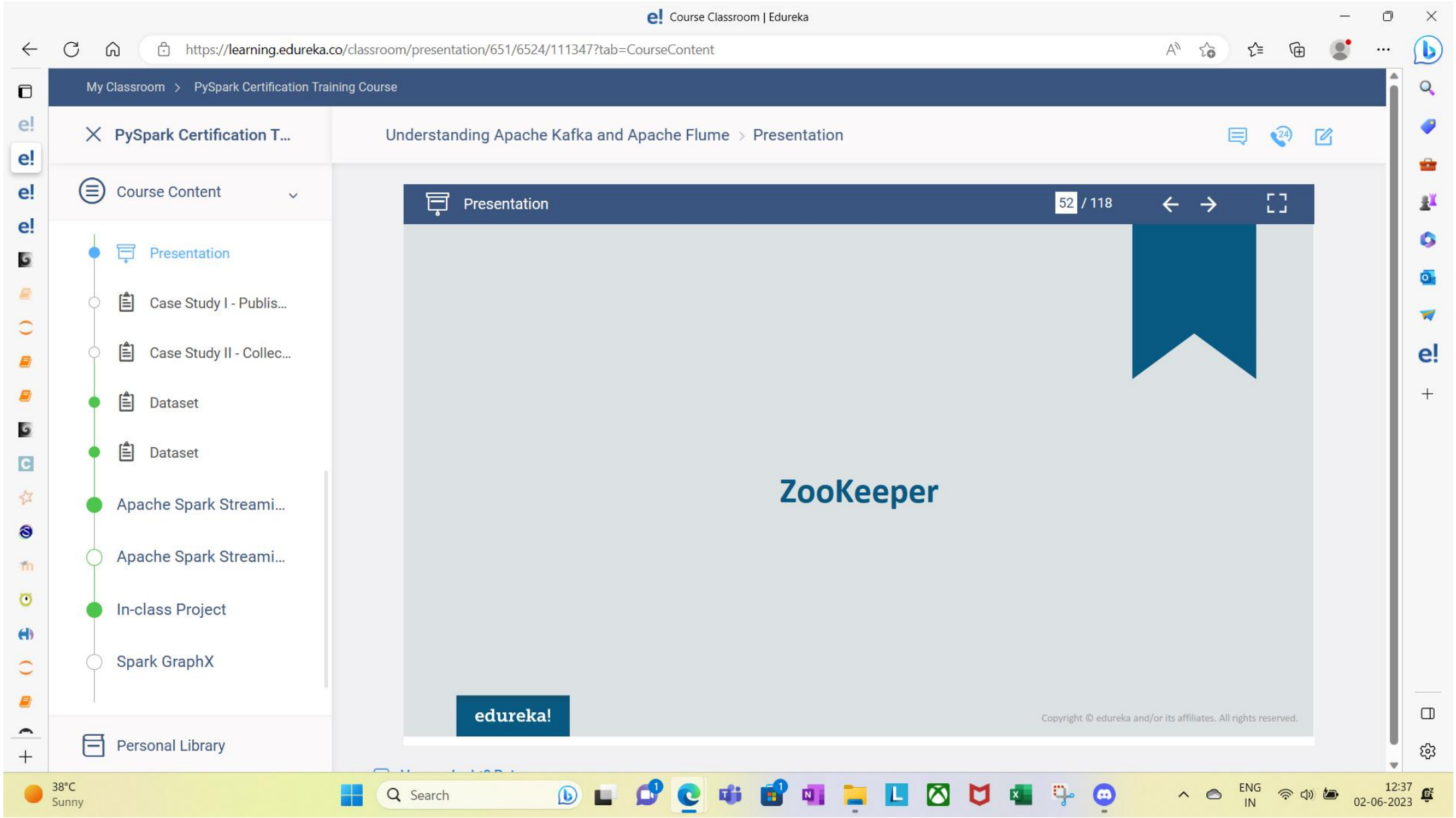
Presentation 51 / 118

Kafka Scale and Speed

How can Kafka scale if multiple producers and consumers read/write to same Kafka Topic log?

- Kafka writes fast: *Sequential writes* to *filesystem* are *fast* (700 MB or more a second)
- Kafka scales writes and reads by **sharding**:
  - *Topic logs into Partitions* (parts of a Topic log).
  - Topics logs can be split into multiple partitions which can be stored on multiple different servers, and those servers can use multiple disks.
  - *Multiple Producers* can *write* to *different Partitions* of the *same Topic*.
  - *Multiple Consumers Groups* can *read* from *different partitions* efficiently.

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
52 / 118
ZooKeeper
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 53 / 118

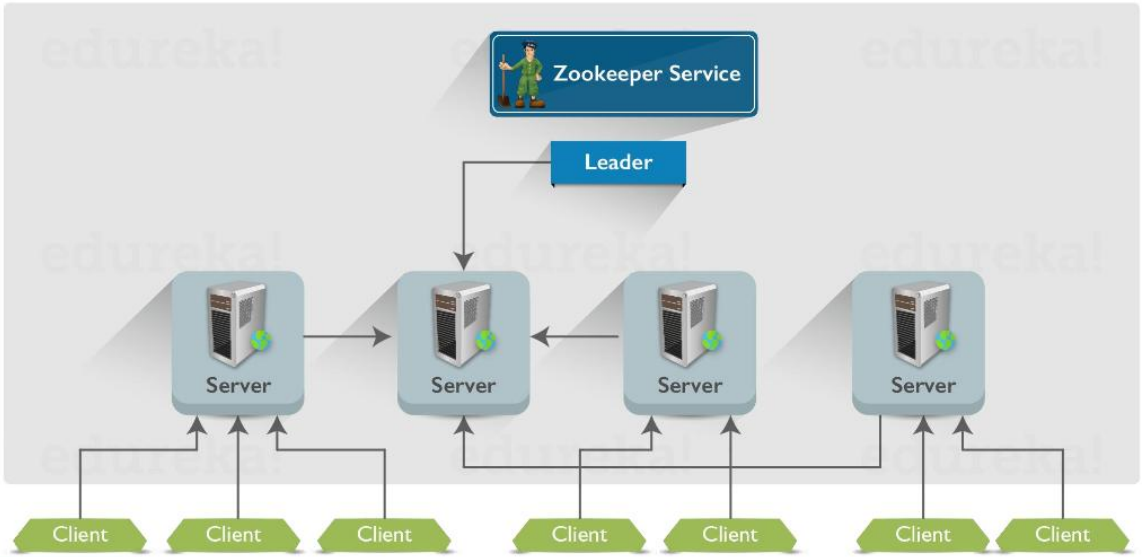
ZooKeeper

- An open source Apache project.
- Provides a *centralized infrastructure* and *services* that enable *synchronization across* a *cluster*.
- *Common objects used across* the large *cluster environments* are maintained in Zookeeper.
- Objects such as configuration, *hierarchical naming space*, etc. are maintained in Zookeeper.
- Zookeeper services are used by *large scale applications* to *coordinate distributed processing* across large clusters.

edureka!

Presentation 54 / 118

ZooKeeper Service



edureka!

Presentation 55 / 118

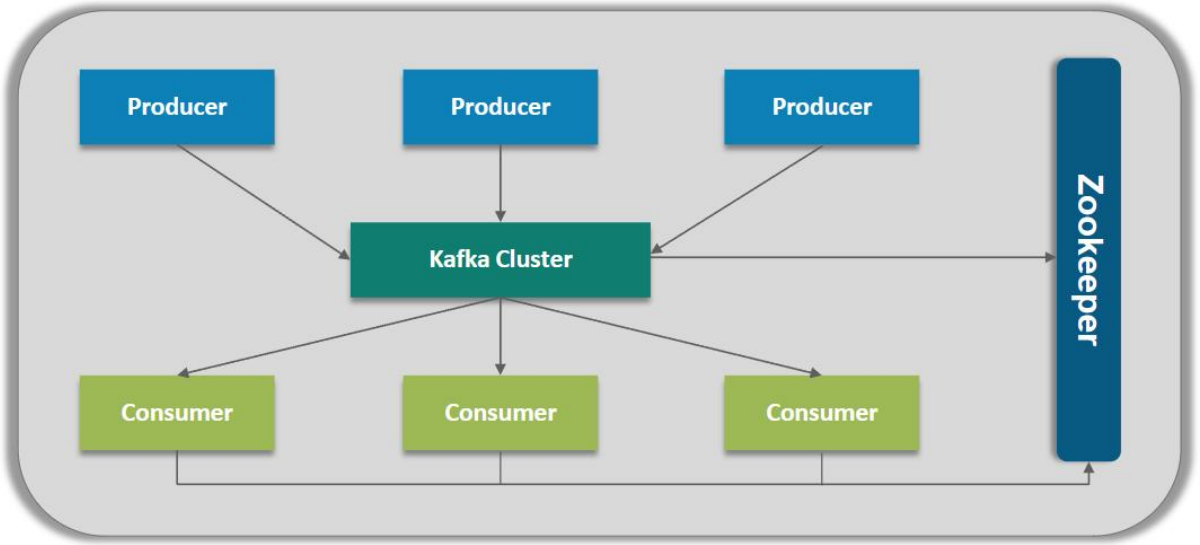
ZooKeeper and Kafka

- Each *Kafka broker coordinates* with *other Kafka brokers using ZooKeeper*.
- *Producers* and *consumers* are *notified* by *ZooKeeper service* about the presence of new broker in Kafka system or failure of the broker in Kafka system.
- Zookeeper is mainly used to *track status of nodes present in Kafka* cluster and also to *keep track of Kafka topics, messages, etc.*

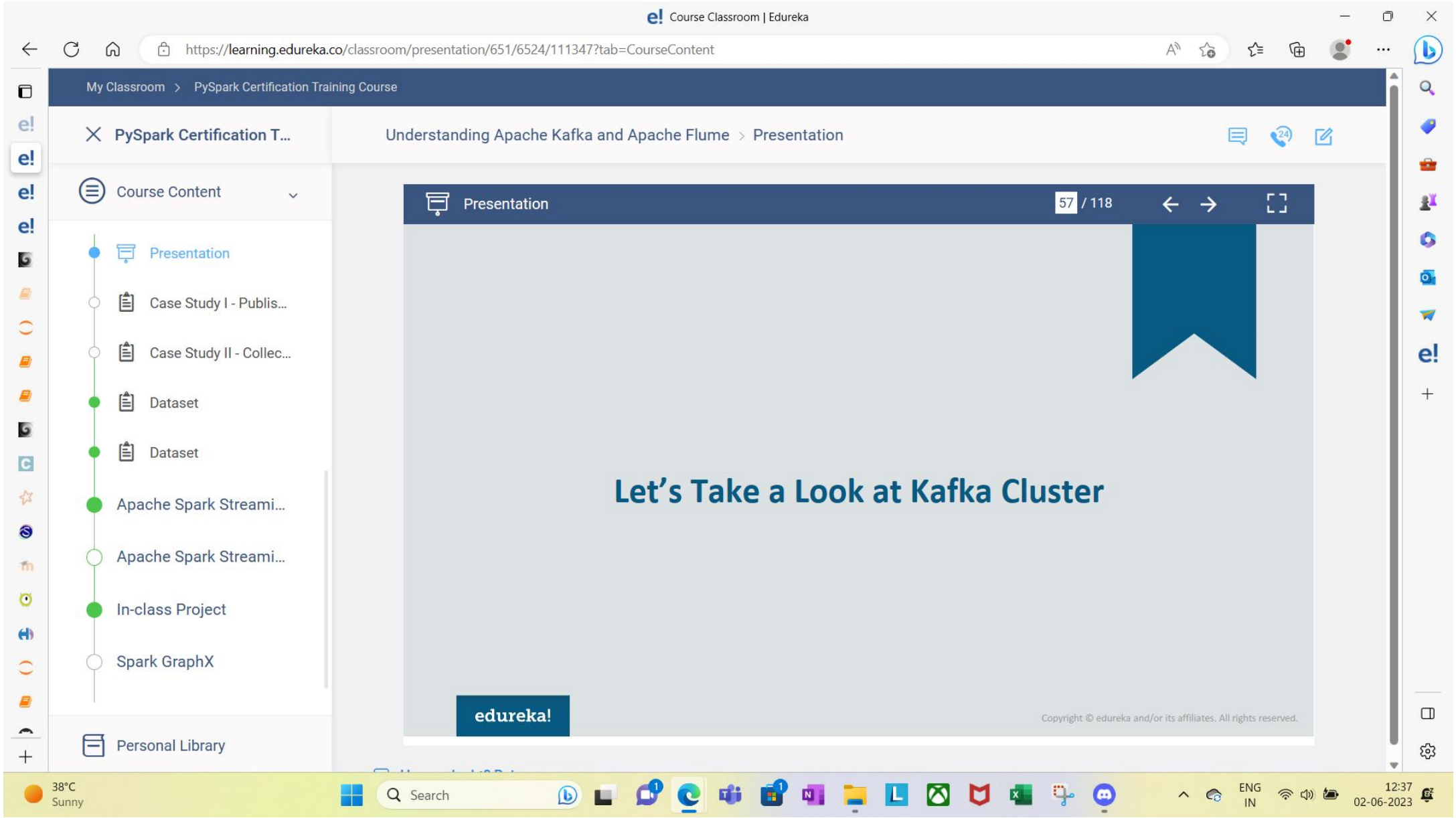
edureka!

Presentation 56 / 118

ZooKeeper and Kafka



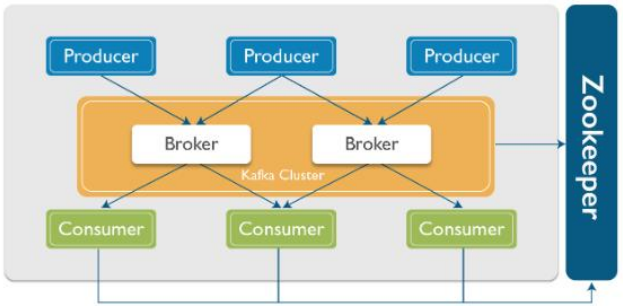
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
57 / 118
Let's Take a Look at Kafka Cluster
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

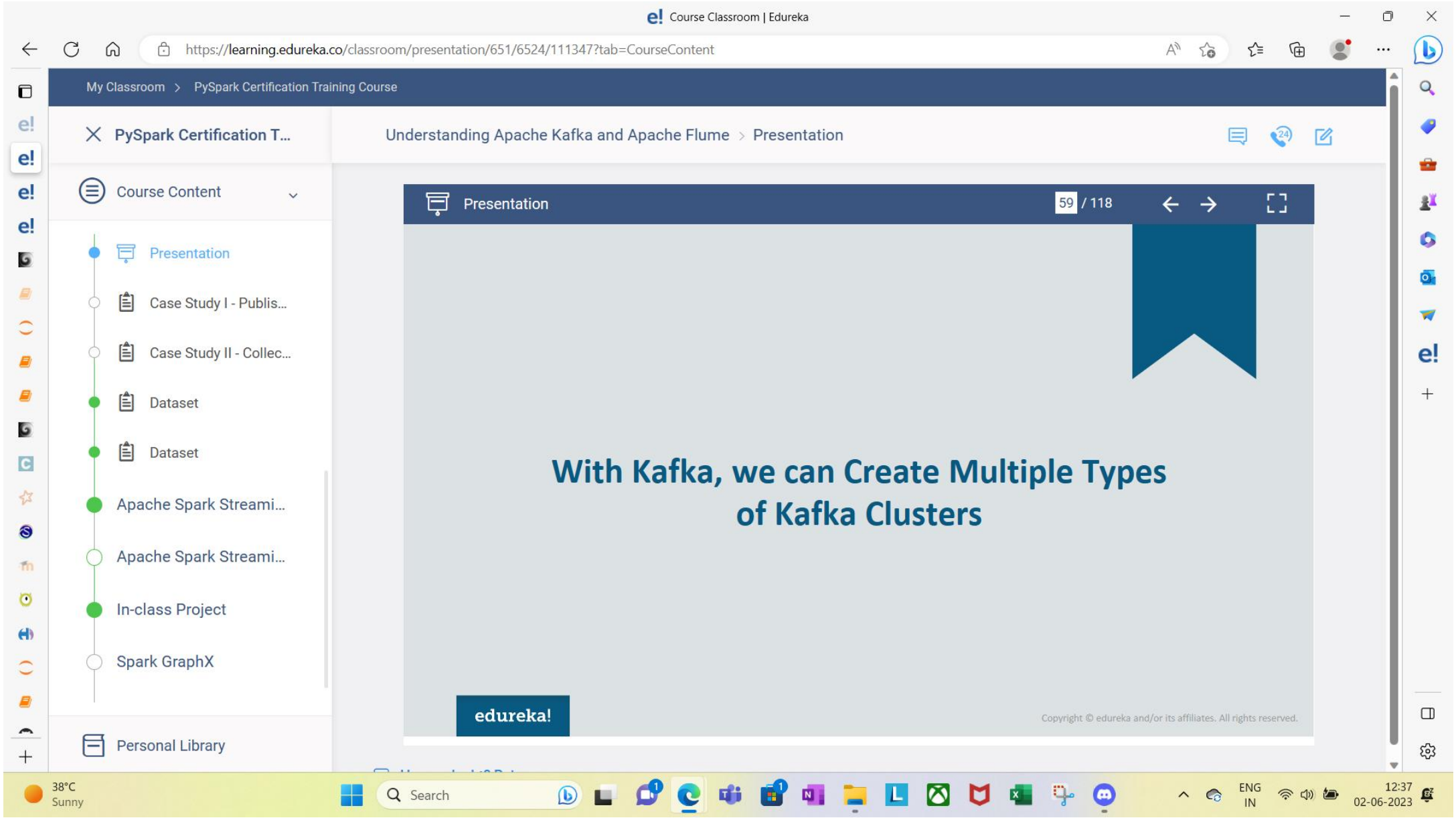
Presentation 58 / 118

Kafka Cluster

- A Kafka Cluster is generally *fast, highly scalable messaging system.*
- A *publish & subscribe messaging system.*
- Can be used effectively in place of Java Messaging System (JMS) and Advanced Messaging Queuing Protocol(AMQP).
- Can be *integrated* with *Hadoop Ecosystem.*
- *Expanding* of the *cluster* can be done with *ease.*
- *Effective* for applications which involves *large scale message processing.*



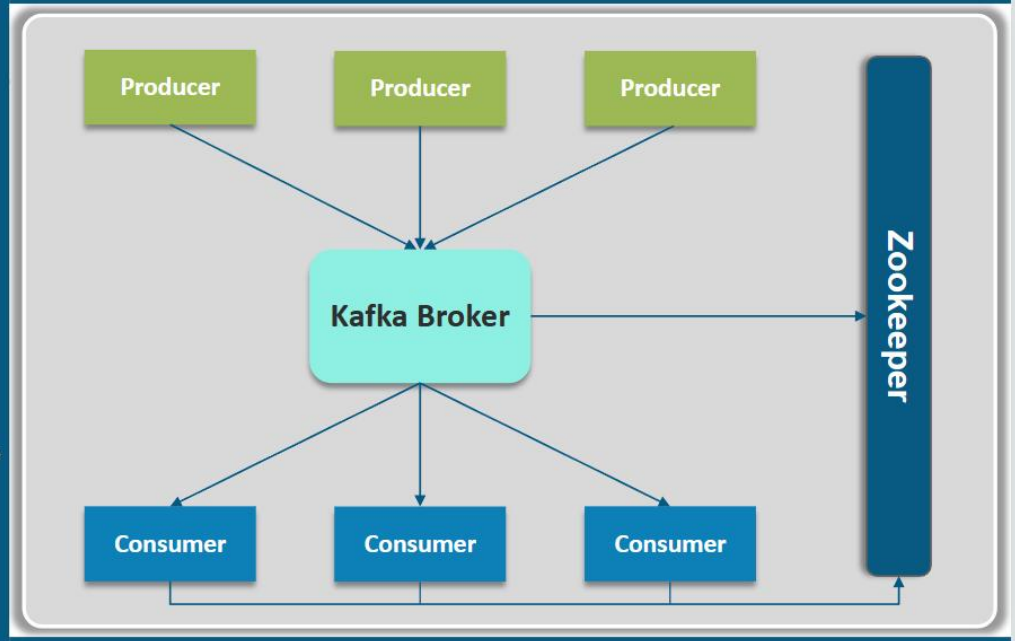
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
59 / 118
With Kafka, we can Create Multiple Types of Kafka Clusters
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 60 / 118

Types of Kafka Clusters

1. Single Node-Single Broker Cluster
2. Single Node-Multiple Broker Cluster
3. Multiple Nodes-Multiple Broker Cluster

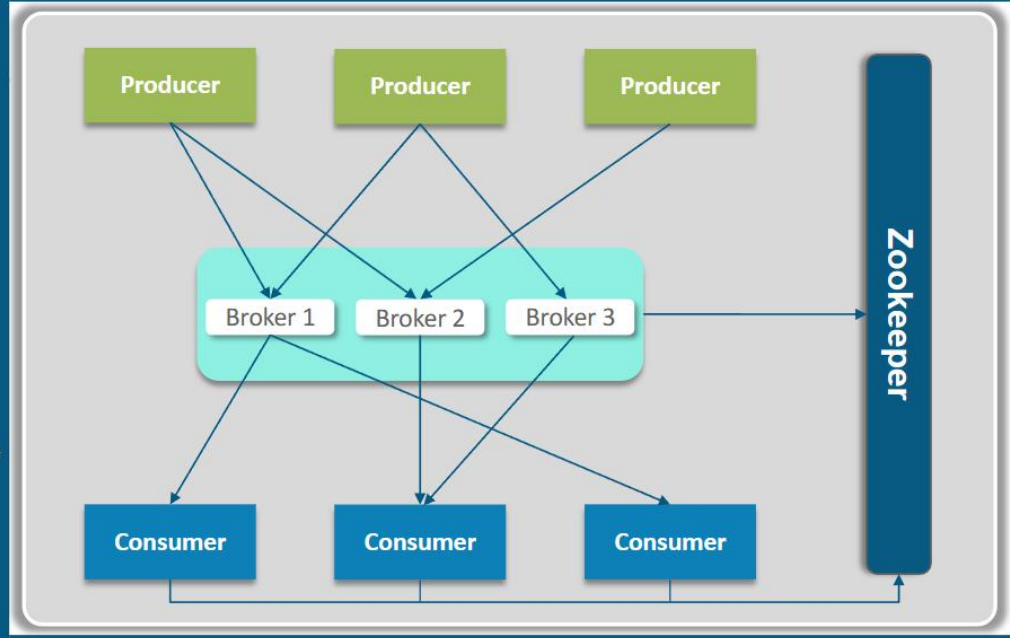


edureka!

Presentation 61 / 118

Types of Kafka Clusters

1. Single Node-Single Broker Cluster
2. **Single Node-Multiple Broker Cluster**
3. Multiple Nodes-Multiple Broker Cluster

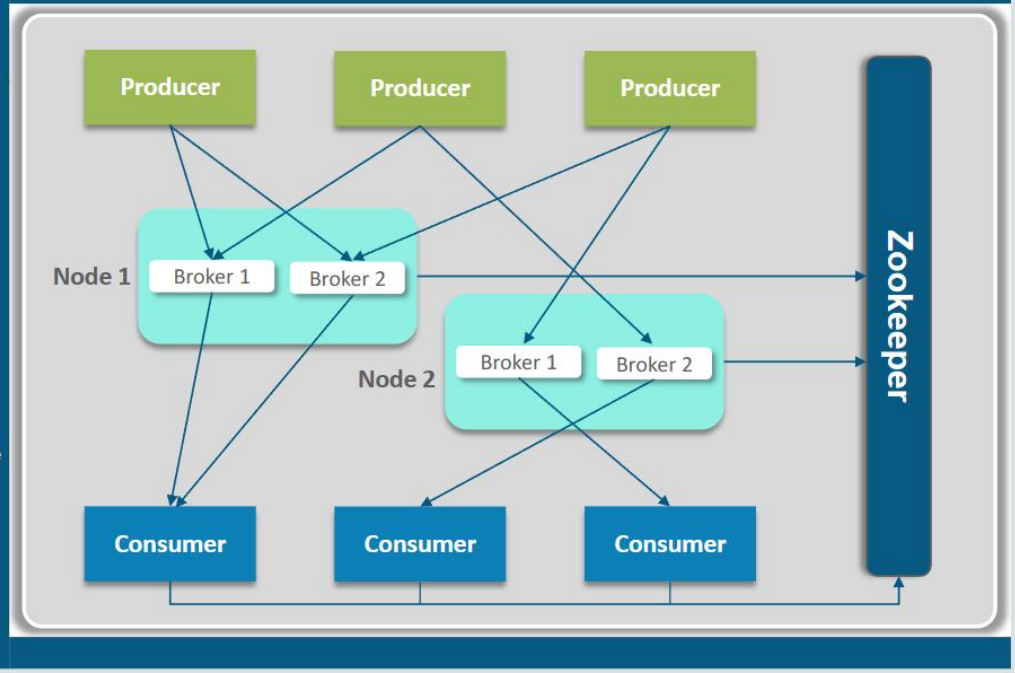


edureka!

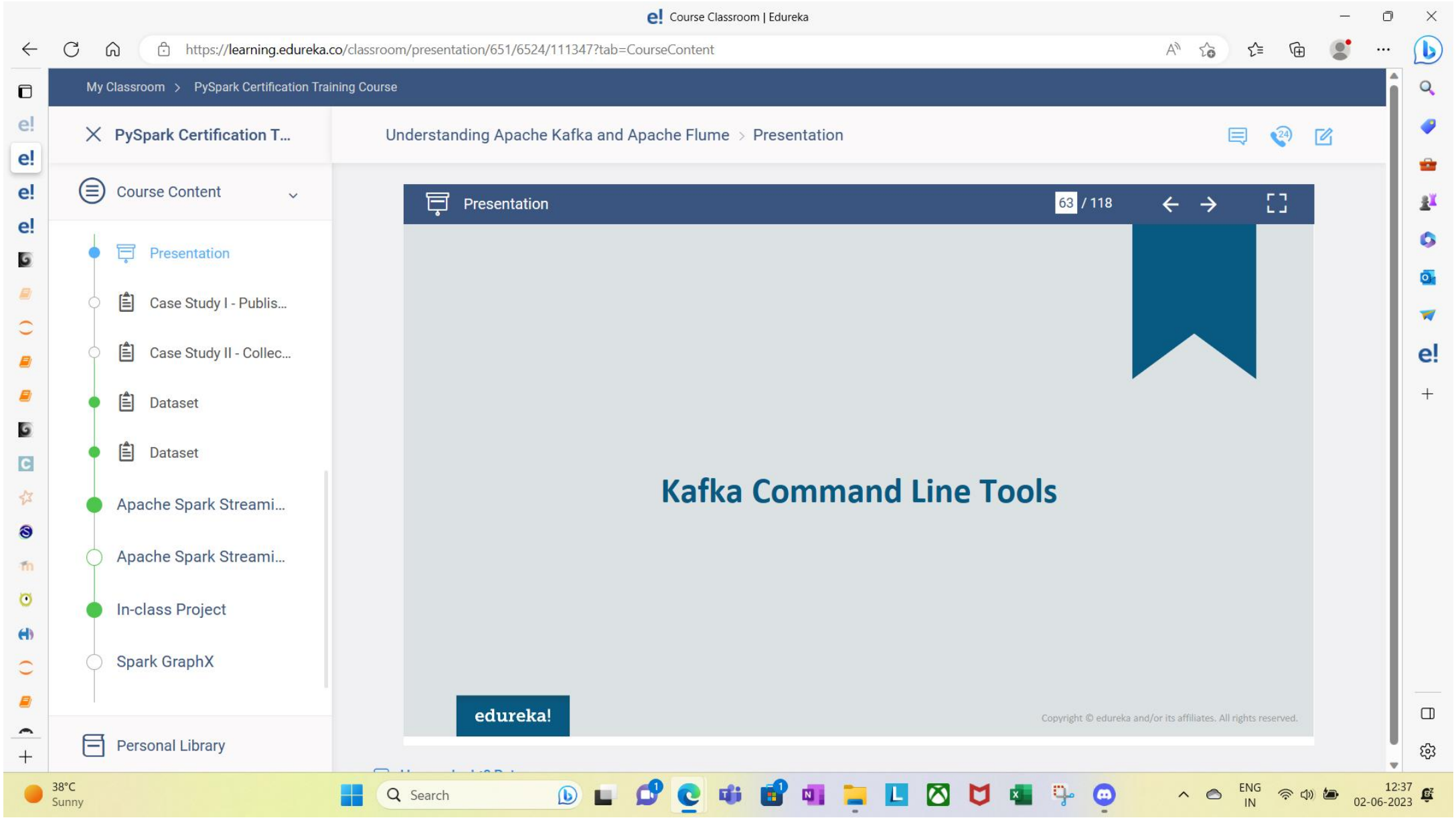
Presentation 62 / 118

Types of Kafka Clusters

1. Single Node-Single Broker Cluster
2. Single Node-Multiple Broker Cluster
3. **Multiple Nodes-Multiple Broker Cluster**



edureka!

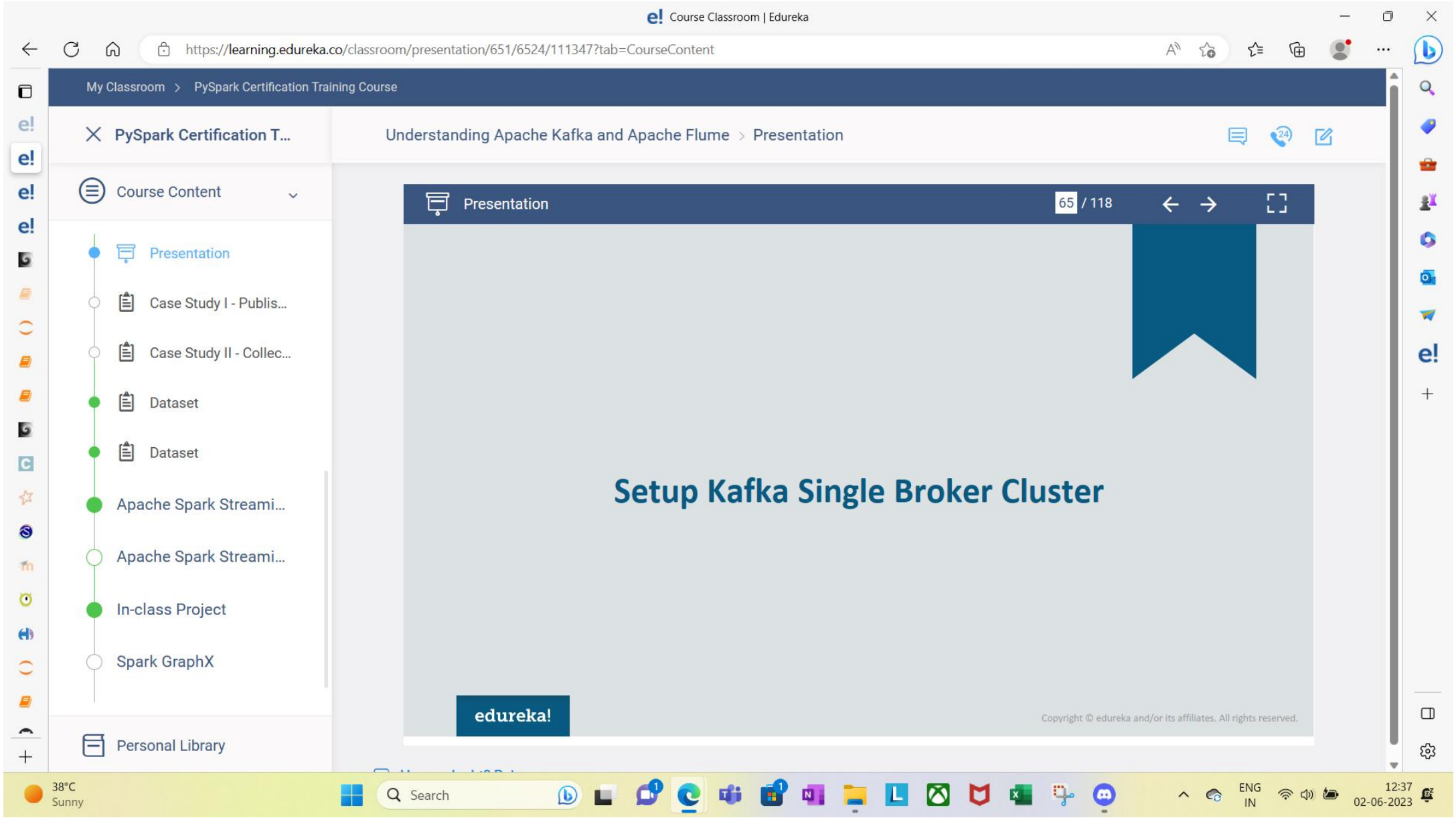

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
63 / 118
Kafka Command Line Tools
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Testing Kafka Cluster

- Kafka Cluster can run against the following broker model:
  - *Single Broker Cluster*
  - *Multi Broker Cluster*
- *Single broker cluster* generally *runs only one instance* compared to *multi broker* which *runs multiple instances*.
- To test the Kafka Cluster the following shell scripts can be used:

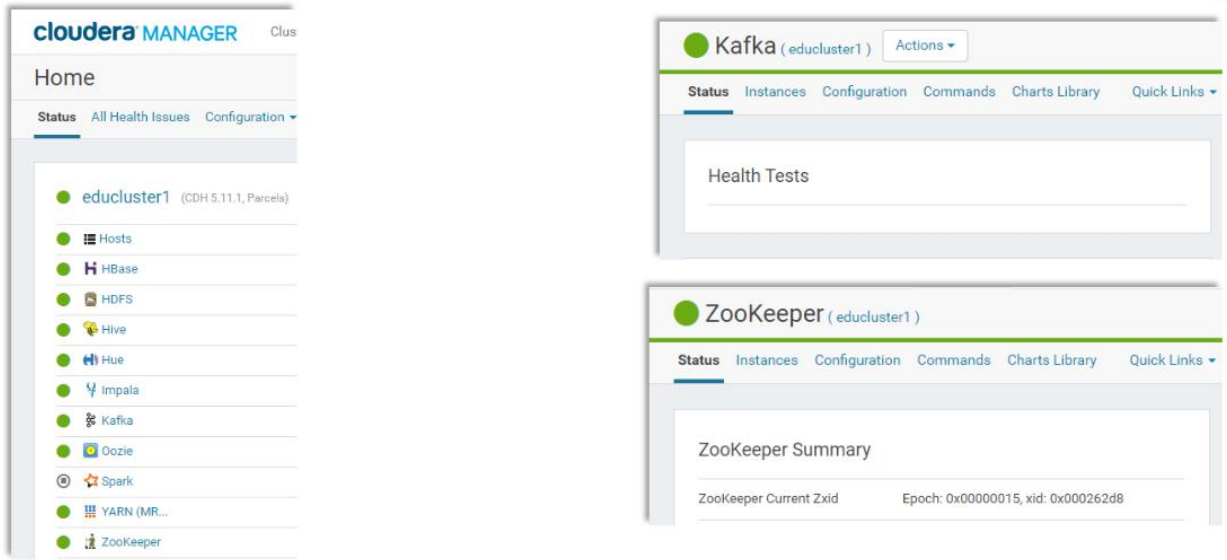
Kafka Shell Scripts	Description
zookeeper-server-start.sh	Starts the zookeeper with the properties configured under config/zookeeper.properties
kafka-server-start.sh	Starts the Kafka server with the properties configured under config/server.properties
kafka-topics.sh	Used to create and list topics
kafka-console-producer.sh	Command line client to send messages to the Kafka Cluster
kafka-console-consumer.sh	Command line client to consume messages from the Kafka Cluster

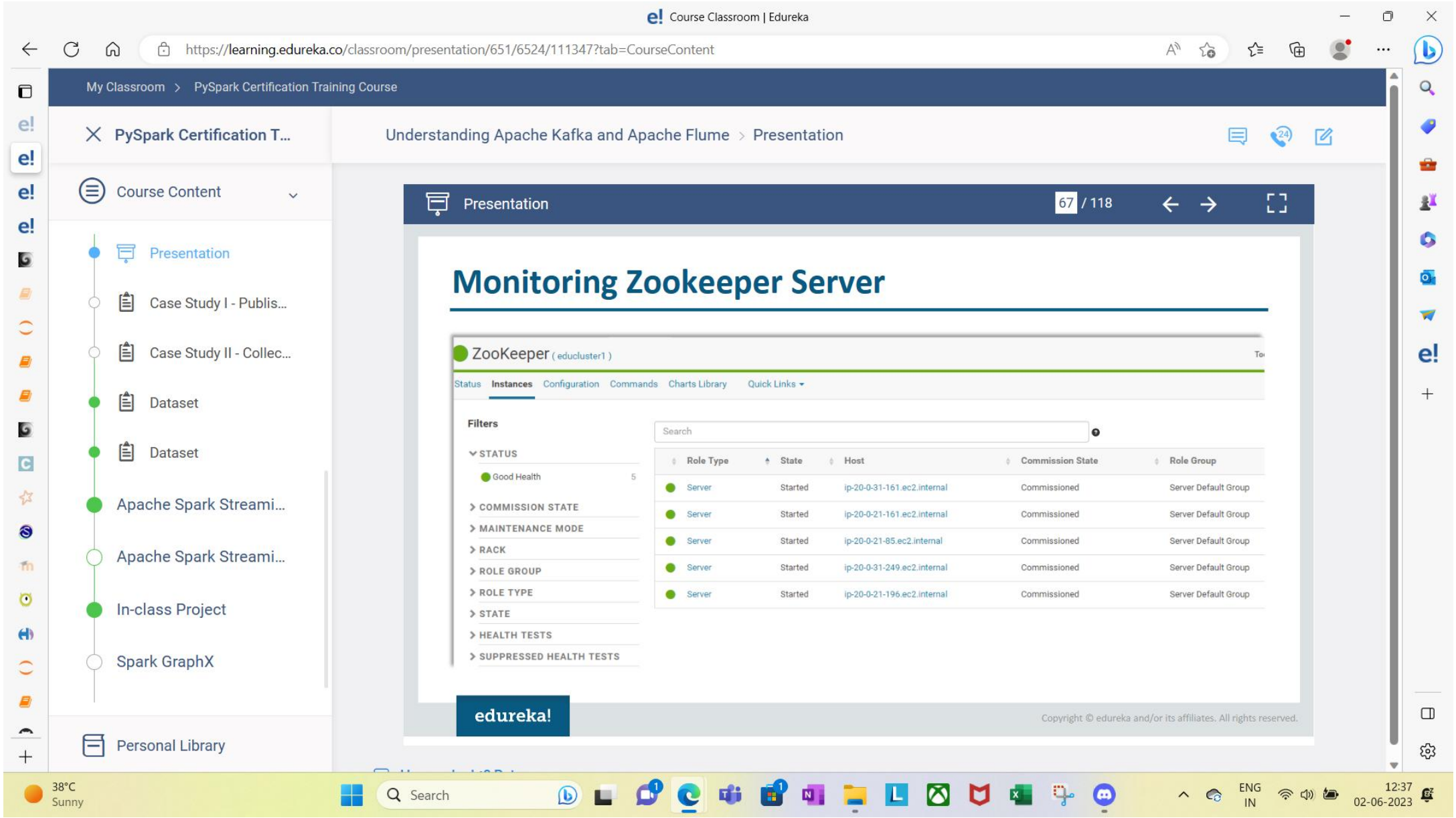
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
65 / 118
Setup Kafka Single Broker Cluster
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 66 / 118

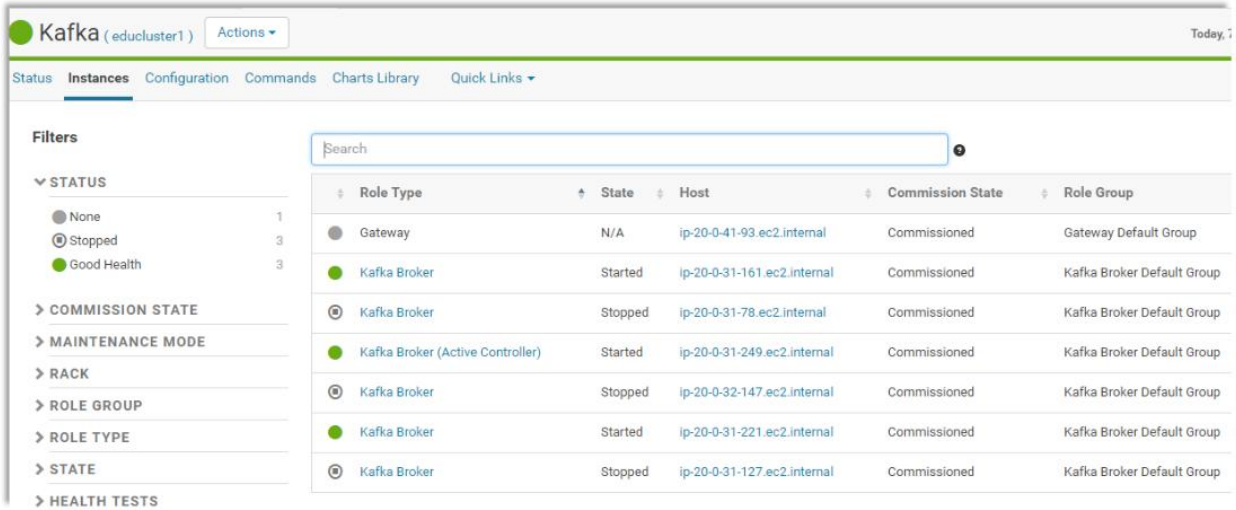
Monitoring Zookeeper & Kafka Server

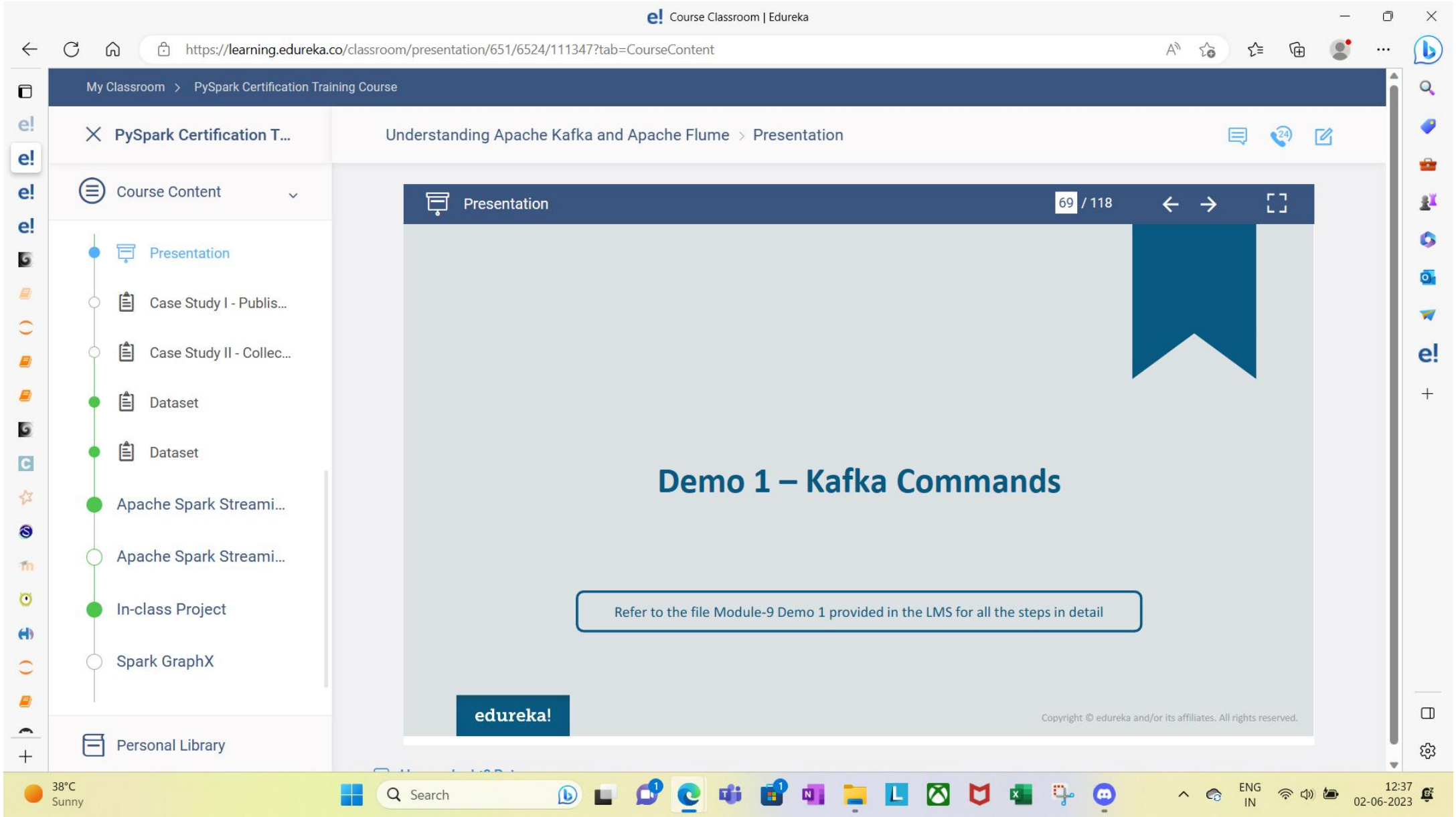


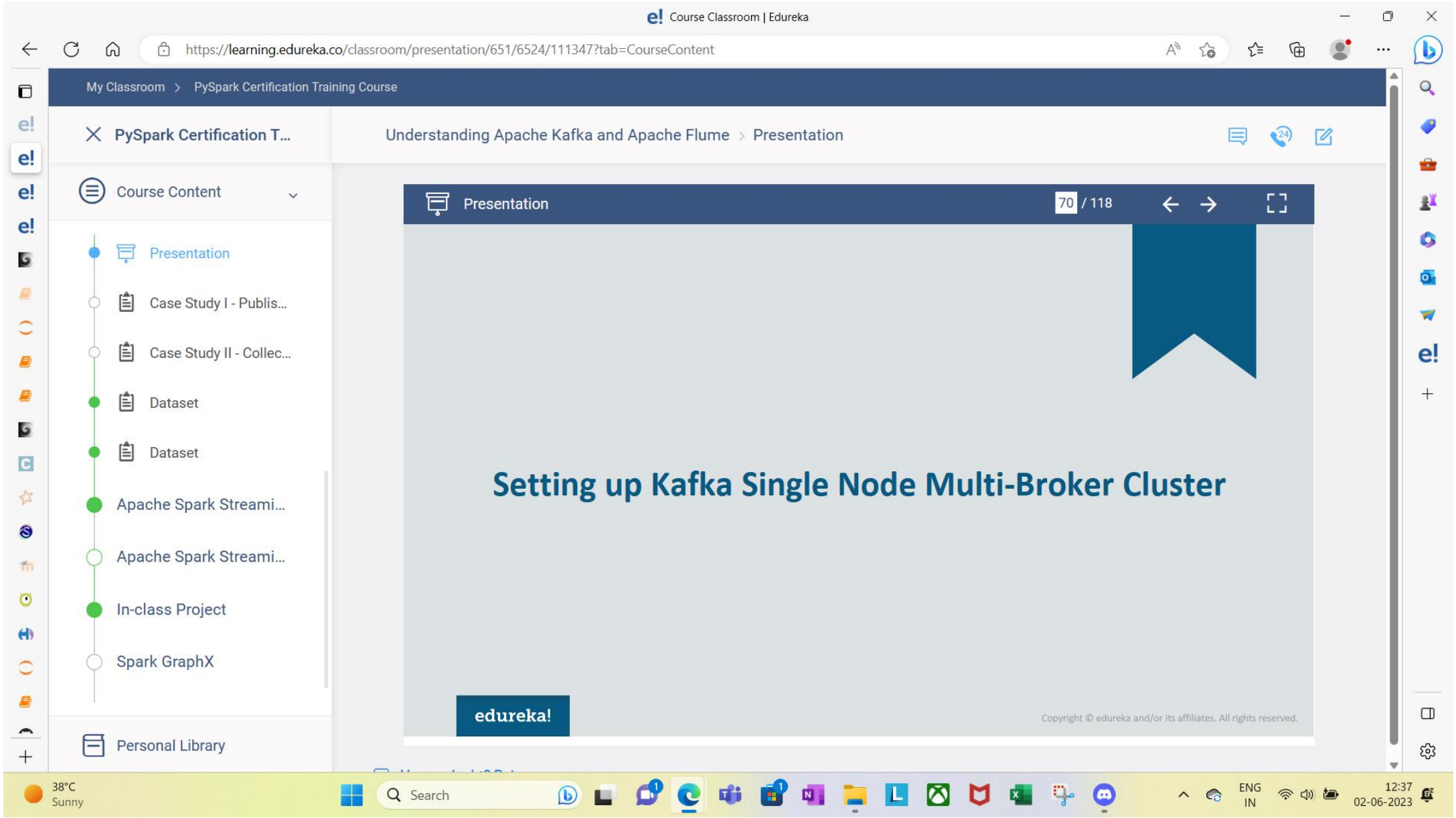

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
67 / 118
Monitoring Zookeeper Server
ZooKeeper (educluster1)
Status Instances Configuration Commands Charts Library Quick Links
Filters
STATUS
Good Health 5
COMMISSION STATE
MAINTENANCE MODE
RACK
ROLE GROUP
ROLE TYPE
STATE
HEALTH TESTS
SUPPRESSED HEALTH TESTS
Search
Role Type State Host Commission State Role Group
Server Started ip-20-0-31-161.ec2.internal Commissioned Server Default Group
Server Started ip-20-0-21-161.ec2.internal Commissioned Server Default Group
Server Started ip-20-0-21-85.ec2.internal Commissioned Server Default Group
Server Started ip-20-0-31-249.ec2.internal Commissioned Server Default Group
Server Started ip-20-0-21-196.ec2.internal Commissioned Server Default Group
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C Sunny
Search
ENG IN
12:37 02-06-2023

Presentation 68 / 118

Monitoring Kafka Server

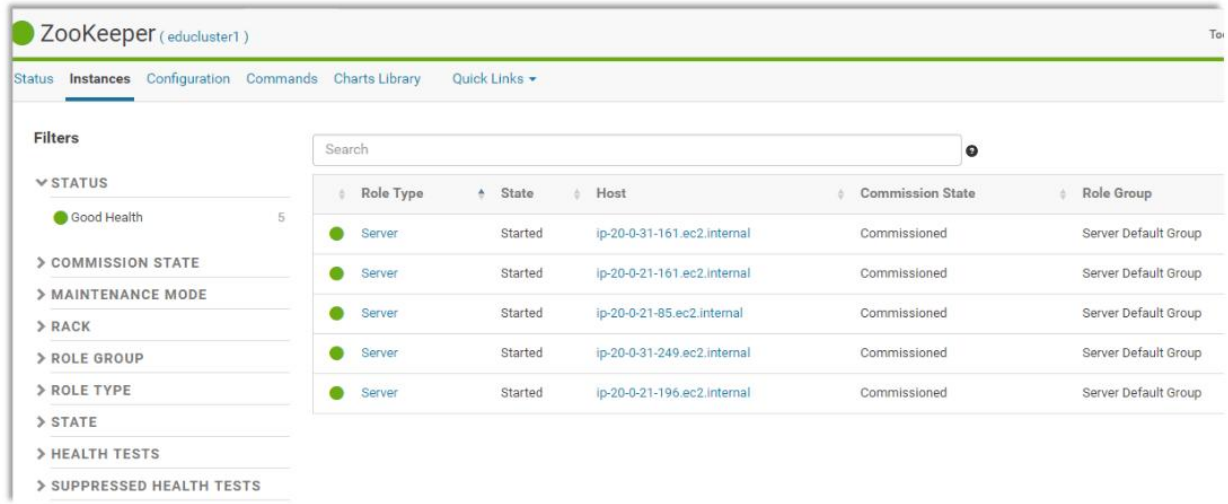



Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
69 / 118
Demo 1 – Kafka Commands
Refer to the file Module-9 Demo 1 provided in the LMS for all the steps in detail
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

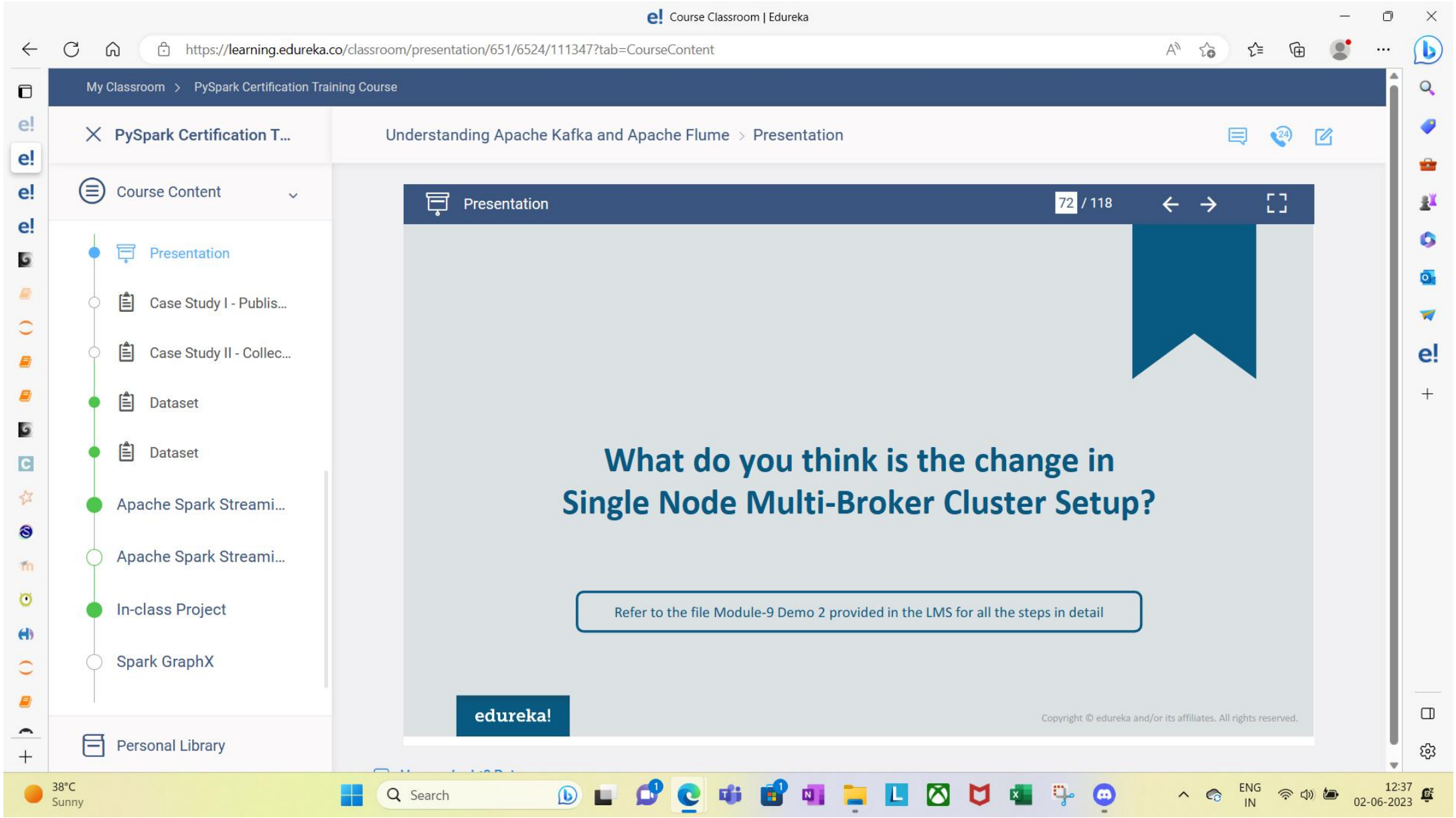

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
70 / 118
Setting up Kafka Single Node Multi-Broker Cluster
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

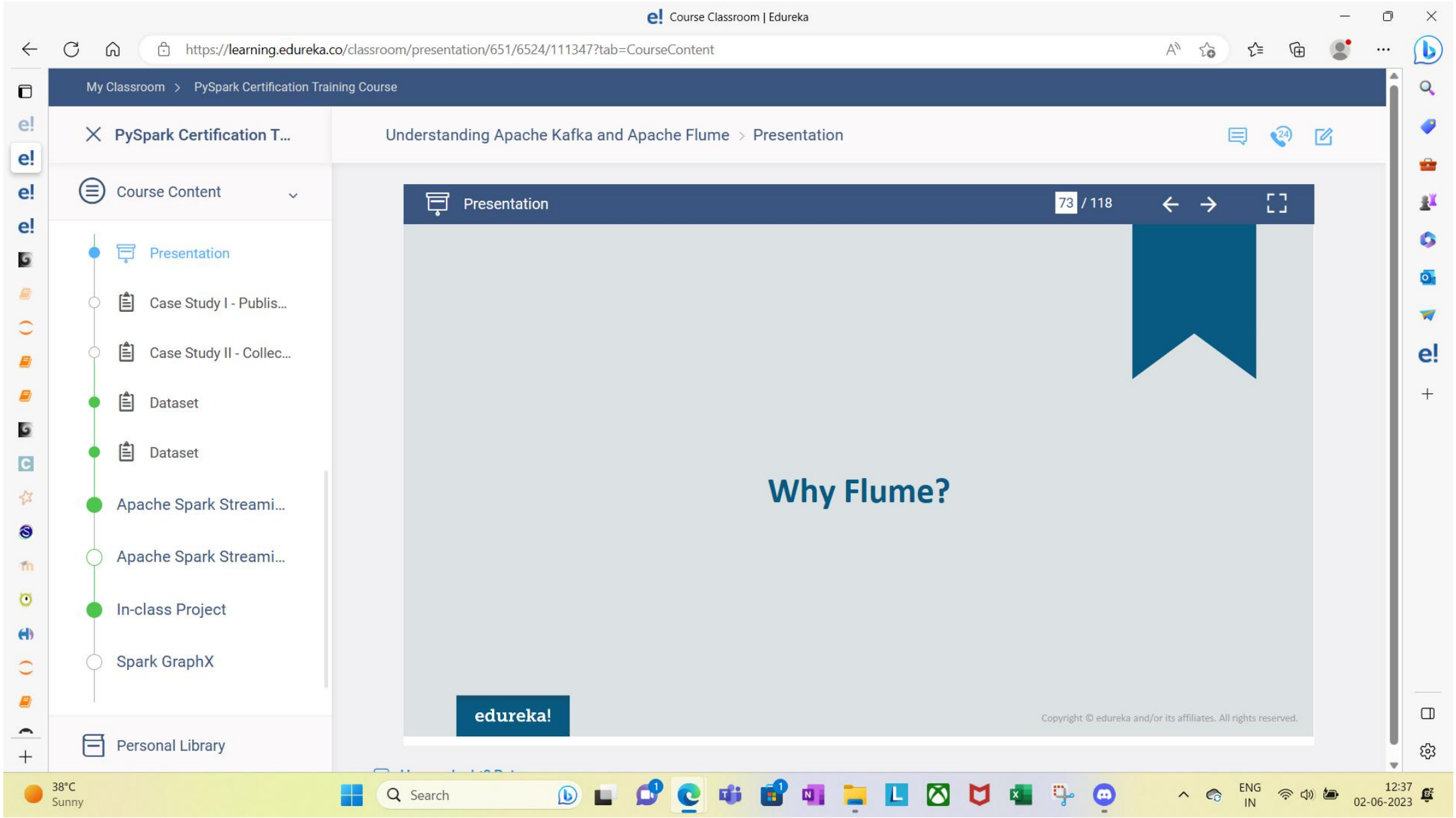
Presentation 71 / 118

Monitor ZooKeeper Server



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
72 / 118
What do you think is the change in Single Node Multi-Broker Cluster Setup?
Refer to the file Module-9 Demo 2 provided in the LMS for all the steps in detail
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
73 / 118
Why Flume?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 74 / 118

Why Flume?

Scenario

Suppose we have hundreds of services running in different servers that produce lots of large logs which should be analysed altogether. We have Hadoop to process them.

Problem

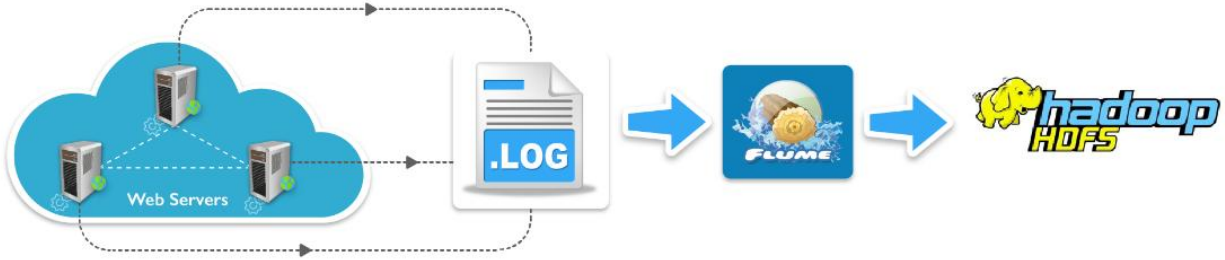
How do we send all the logs to a place that has Hadoop? We need a reliable, scalable, extensible and manageable way to do it!

edureka!

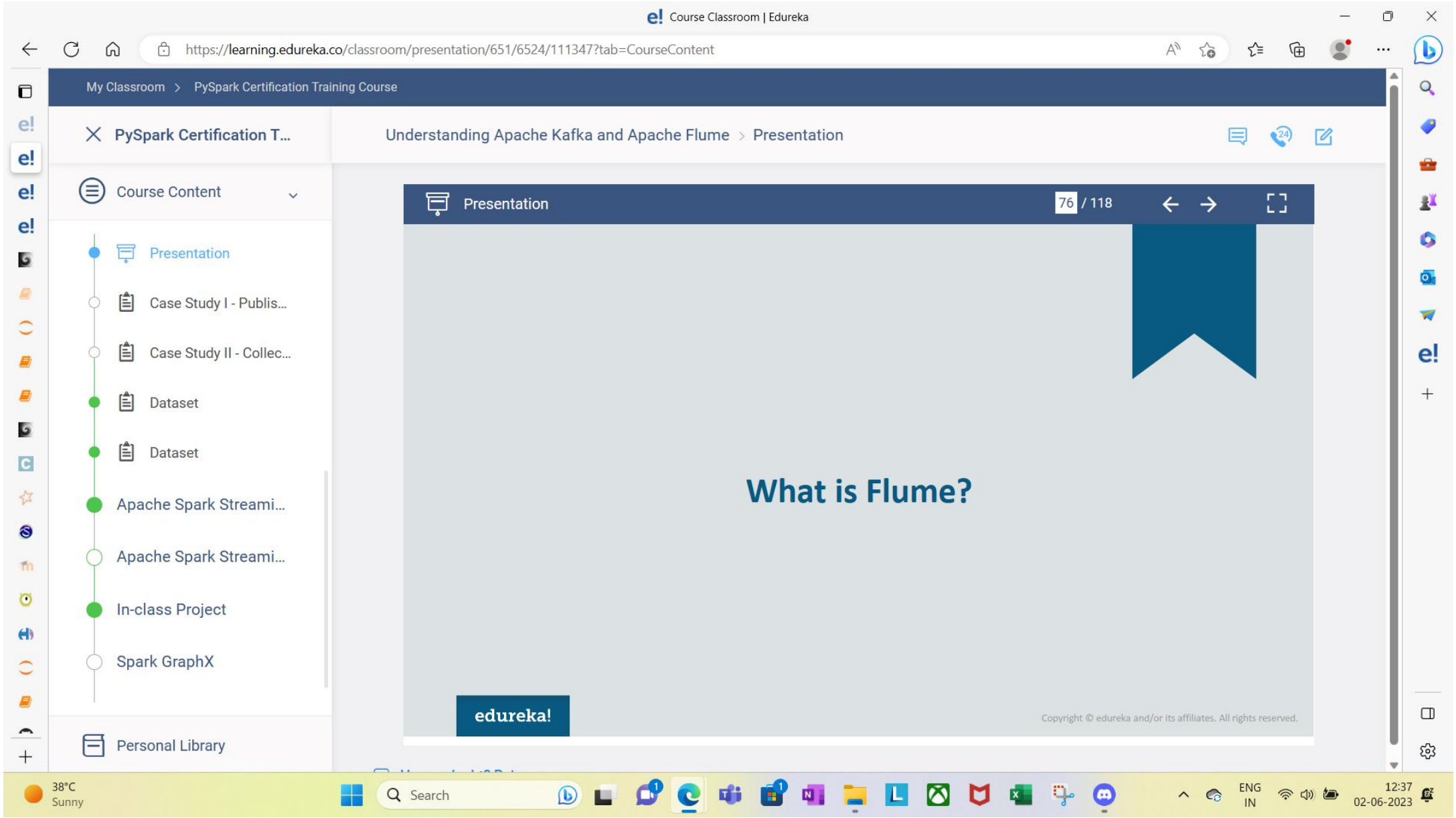
Presentation 75 / 118

Why Flume?

Scenario	Problem	Solution
Suppose we have hundreds of services running in different servers that produce lots of large logs which should be analysed altogether. We have Hadoop to process them.	How do we send all the logs to a place that has Hadoop? We need a reliable, scalable, extensible and manageable way to do it!	**Apache Flume** is the most reliable, distributed, and available service for systematically collecting, aggregating, and moving large amounts of streaming data (logs) into the HDFS.



edureka!

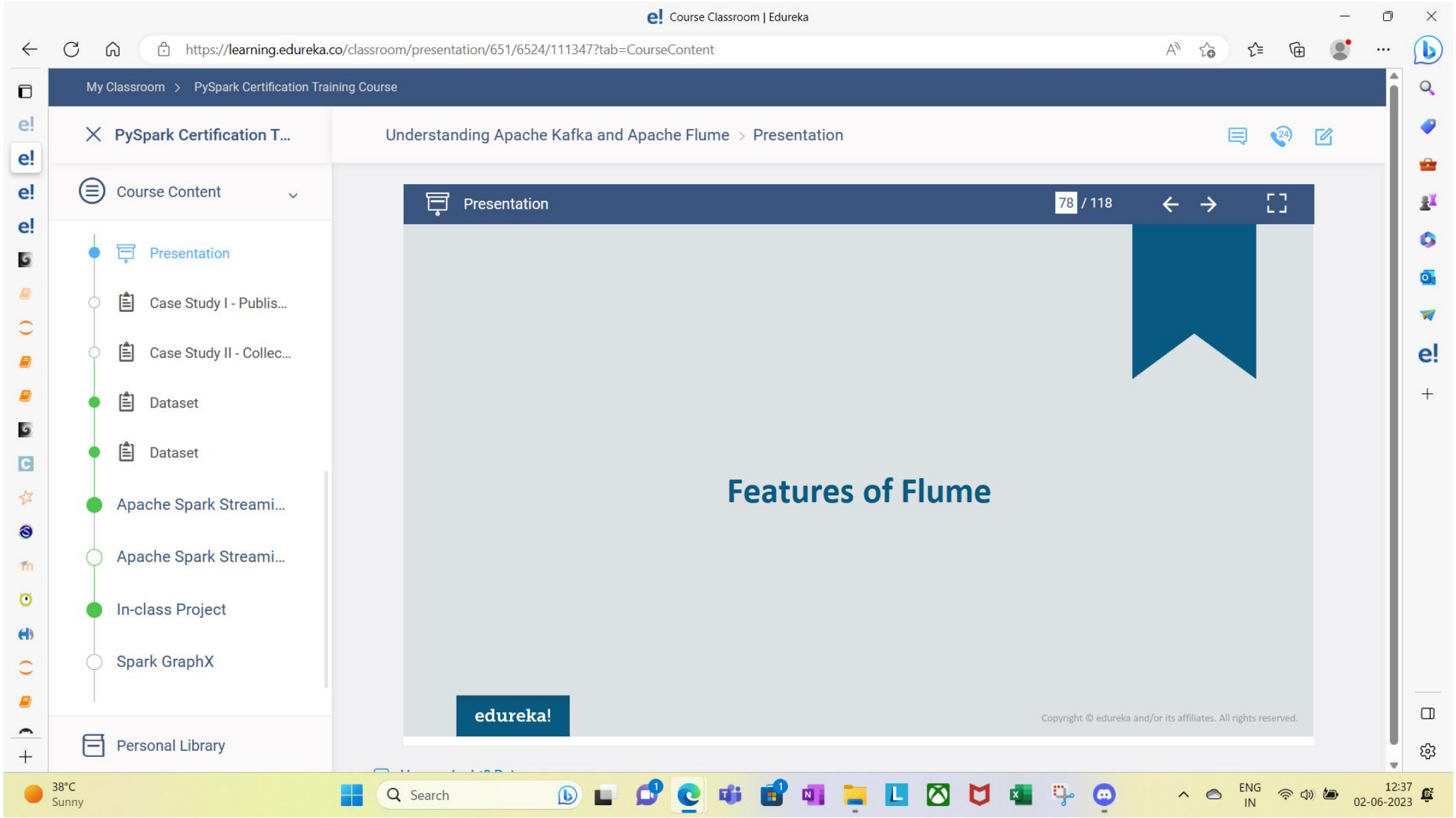

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
76 / 118
What is Flume?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 77 / 118

What is Flume?

"**Flume** is a distributed, reliable and available service for efficiently collecting, aggregating, and moving large amounts of log data from various sources to a centralized data store."

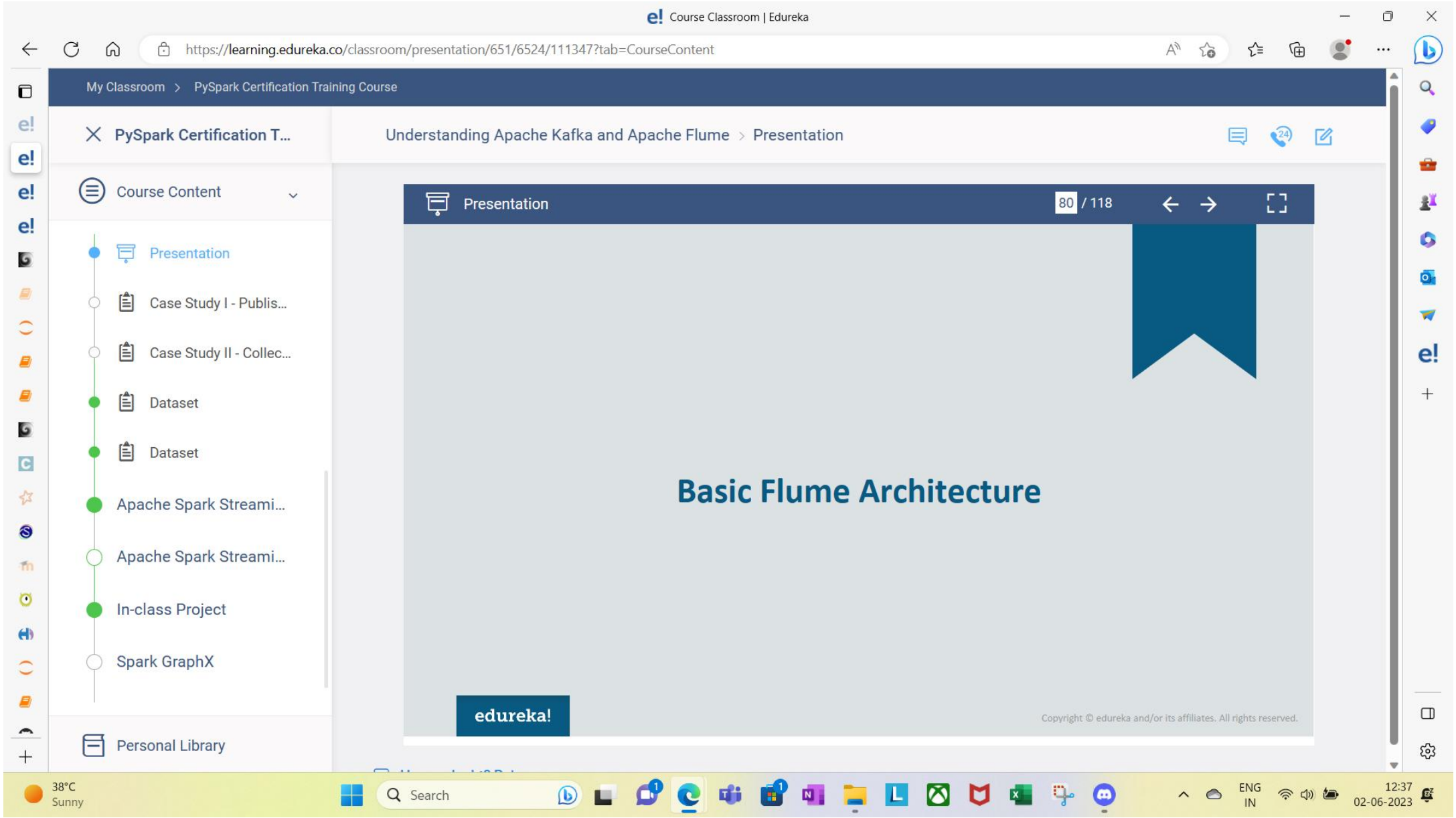
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
78 / 118
Features of Flume
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 79 / 118

Flume Features

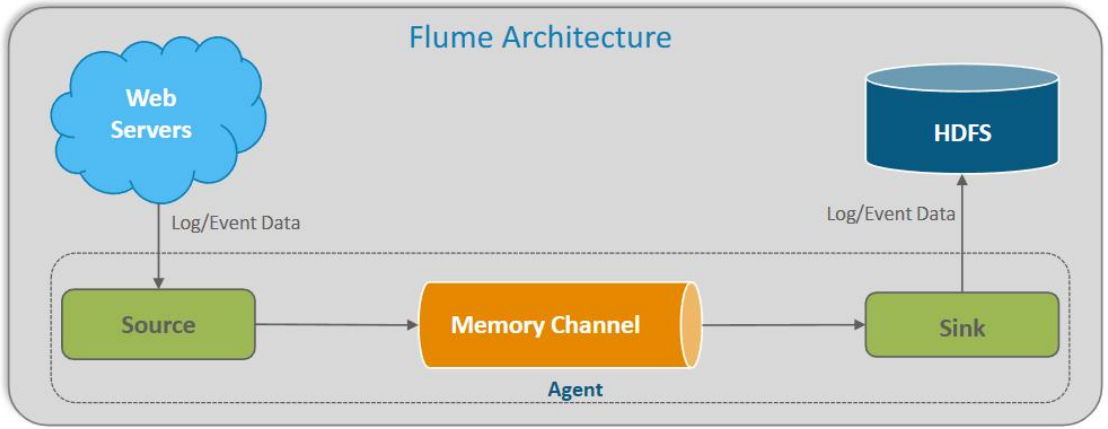
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
80 / 118
Basic Flume Architecture
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

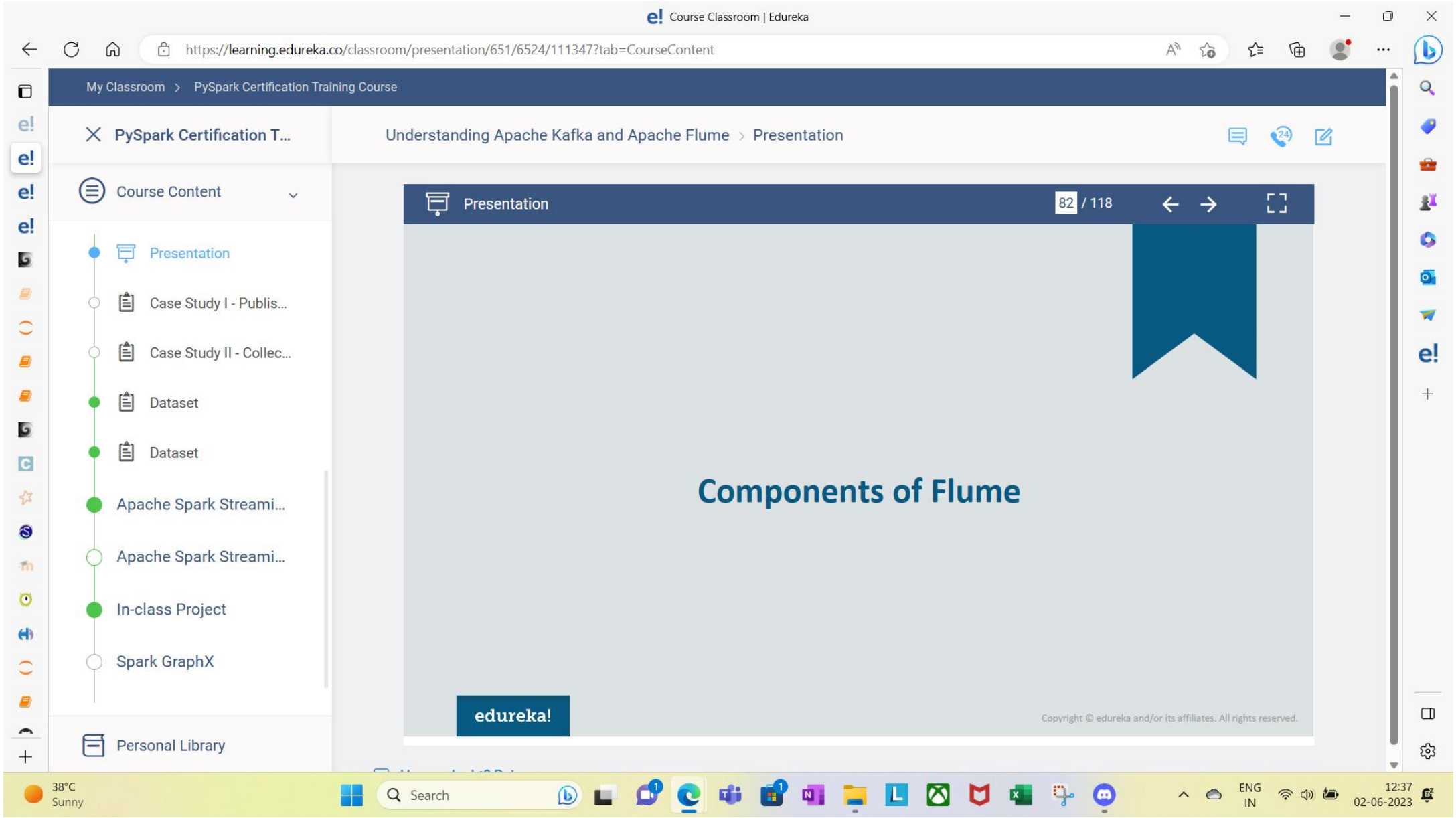
Presentation 81 / 118

Flume Architecture

- **Flume** has a simple and flexible architecture based on streaming **data flow model**.
- It is **robust** and **fault tolerant** with tunable reliability mechanisms and many failover and recovery mechanisms.



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
82 / 118
Components of Flume
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 83 / 118

Flume Components

The entity through which data enters into Flume.
Flume supports Avro, Netcat, Thrift, exec, syslog as source of data.

The conduit between the Source and the Sink.
Sources ingest events into the channel and the sinks drain the channel.

The entity that receive events from the channel.
A variety of sinks allow data to be streamed to a range of destinations. One example is the HDFS sink.

The entity that produces and transmits the event to the source operating within the Agent.

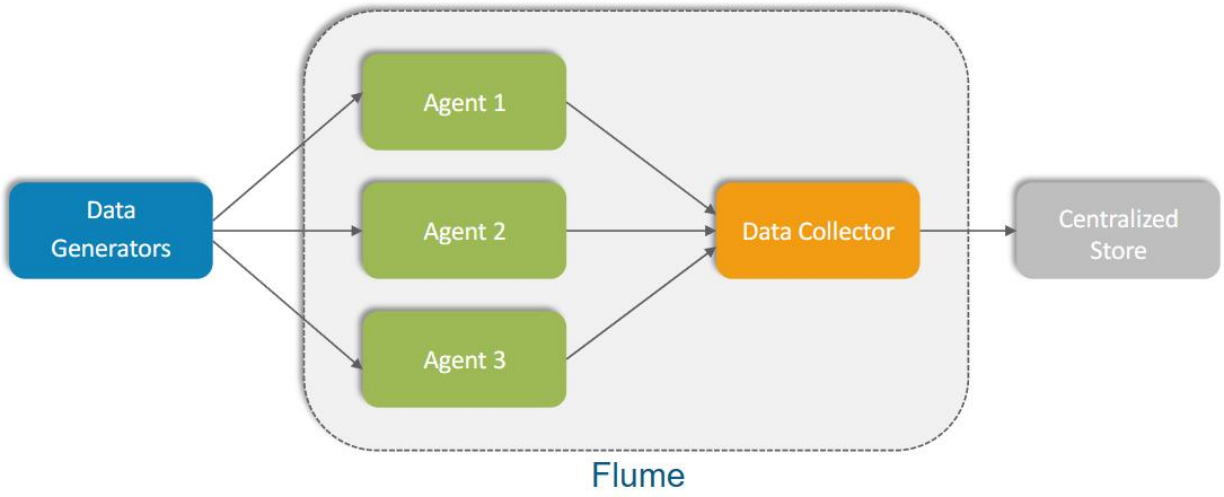
Any physical Java virtual machine running Flume.
It is a collection of sources, sinks and channels.

A singular unit of data that is transported by Flume (typically a single log entry).

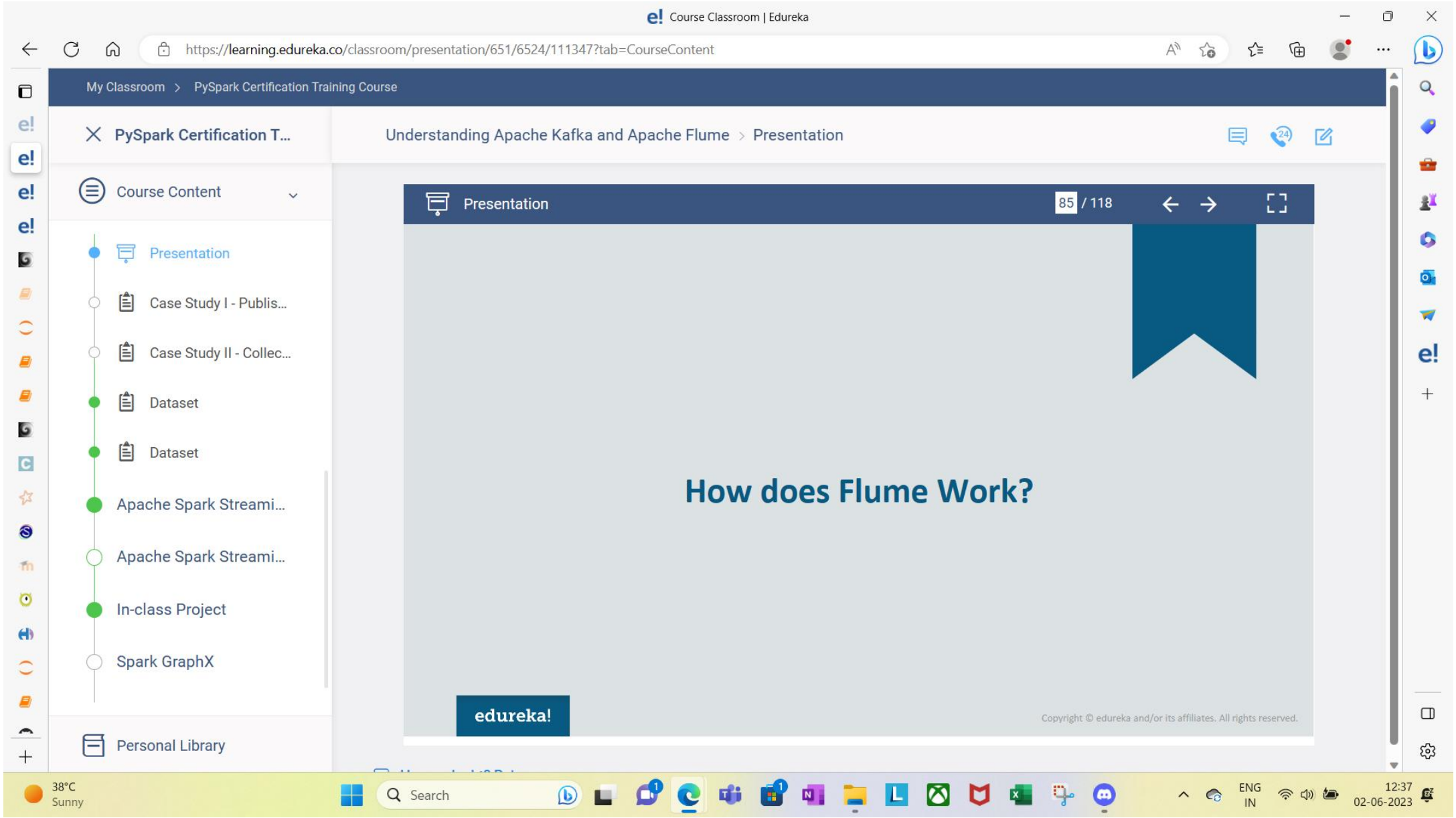
edureka!

Presentation 84 / 118

Flume Architecture : Deep Dive



edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
85 / 118
How does Flume Work?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

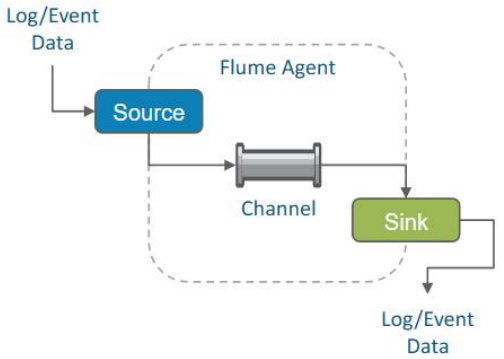
How does Flume Work?

Flume uses agents which have:

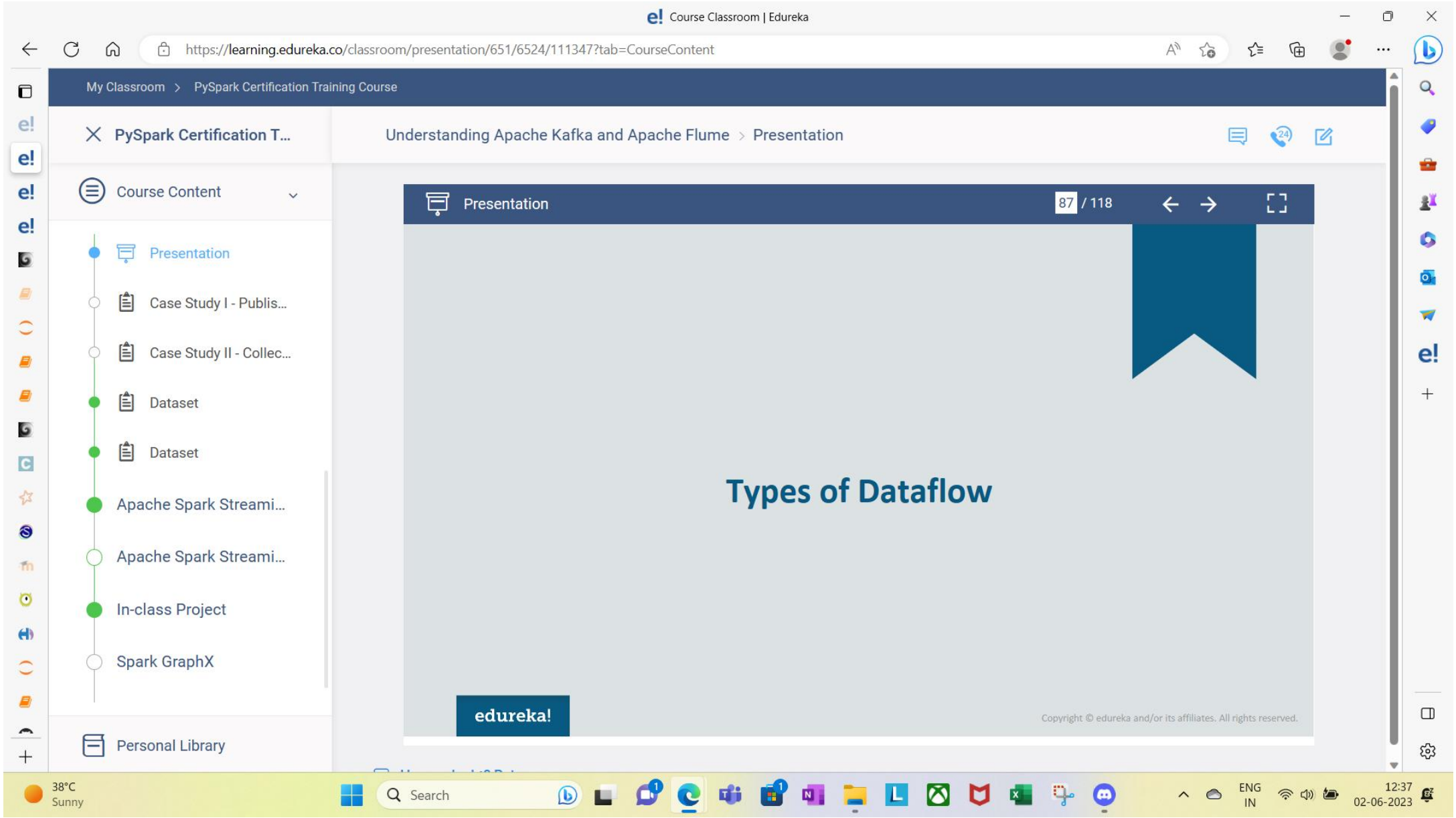
A source
- Listen for events delivered to it by an external source like a web server.
- Write events to channel

A channel
- Channel keeps the event until it's consumed by a Flume sink
- Queue event data as transactions

A sink
- Write event data to target like HDFS
- Remove event from queue



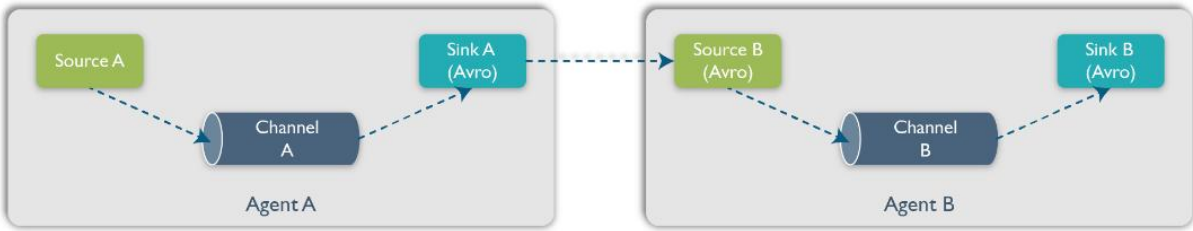
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
87 / 118
Types of Dataflow
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:37
02-06-2023

Presentation 88 / 118

Multi-Agent Flow

Within Flume, there can be **multiple agents** and before reaching the final destination, an event may travel through more than one agent. This is known as **multi-agent flow**.

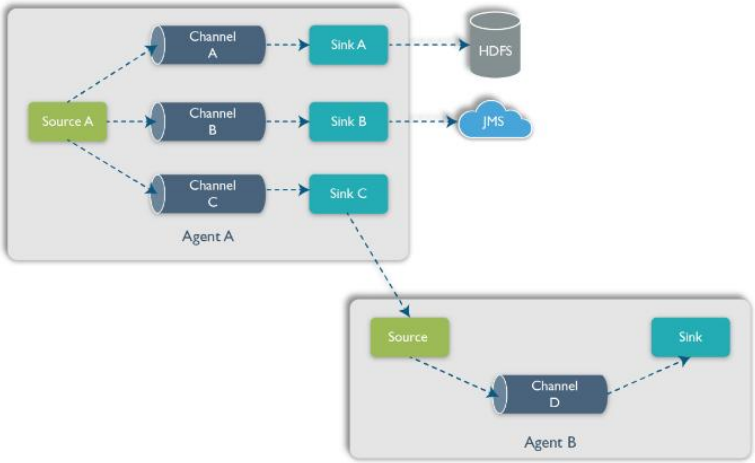


edureka!

Presentation 89 / 118

Fan-Out Flow : Multiplexing the Flow

- The dataflow from one source to multiple channels is known as **fan-out flow**.
- Flume supports multiplexing the event flow to one or more destinations.
- This is achieved by defining a flow multiplexer that can replicate or selectively route an event to one or more channels.

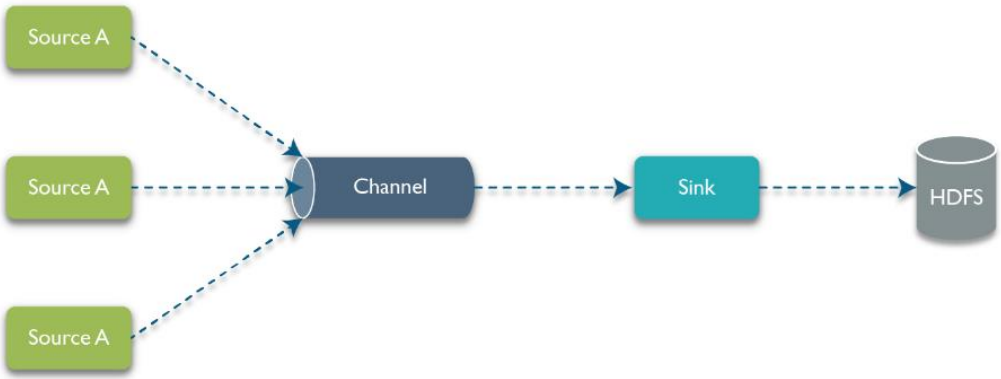


edureka!

Presentation 90 / 118

Fan-In Flow

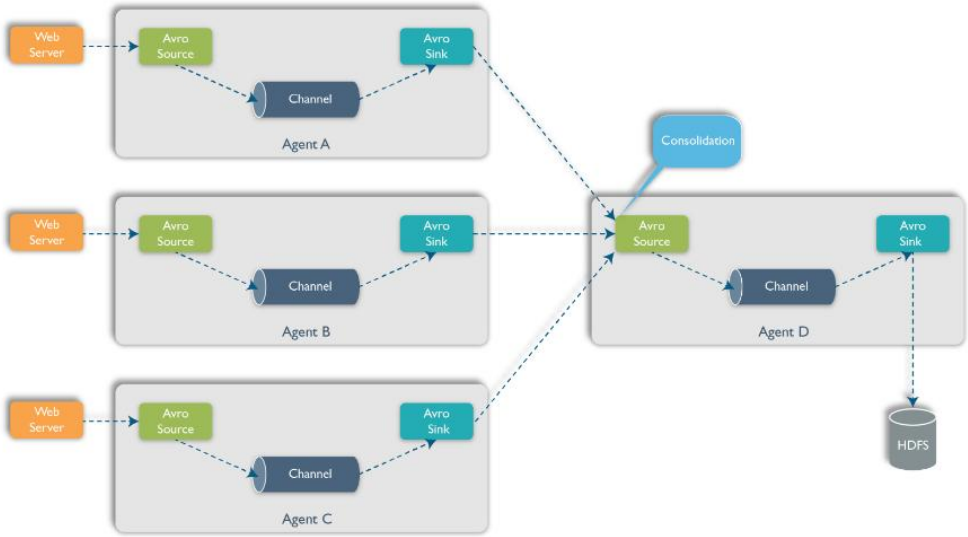
The data flow in which the data will be transferred from many sources to one channel is known as **fan-in flow**.



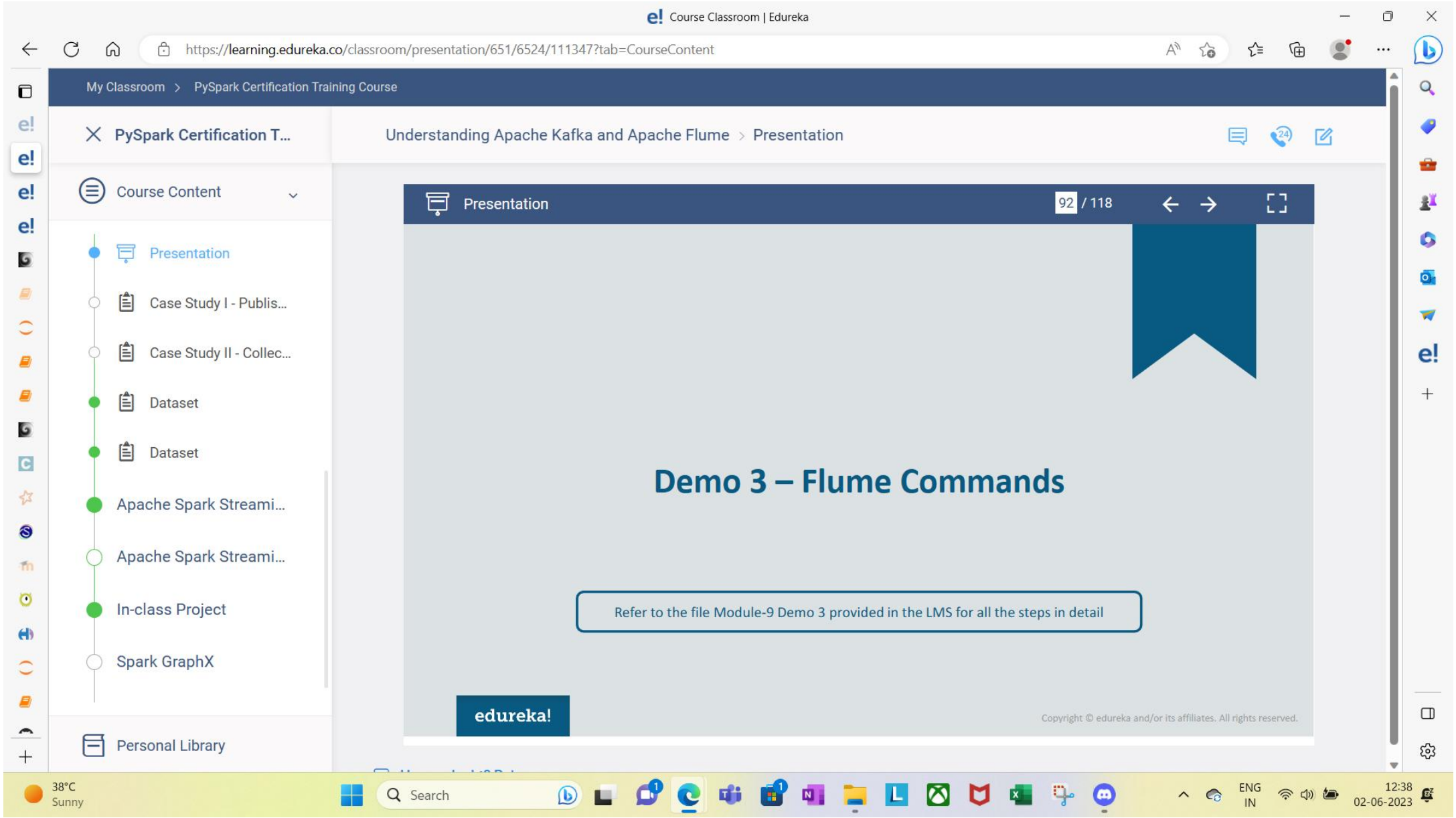
edureka!

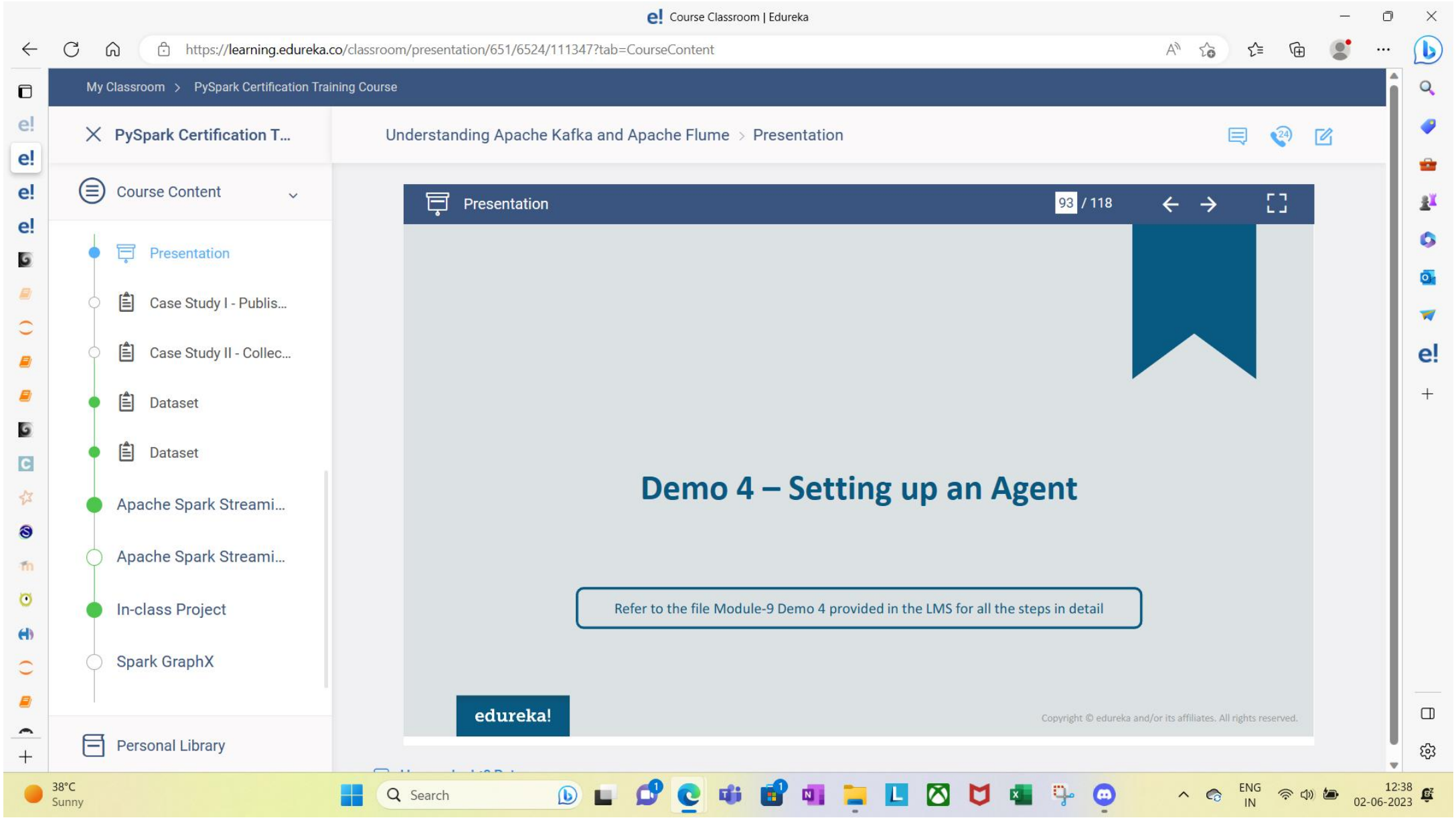
Presentation 91 / 118

Consolidation



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
92 / 118
Demo 3 – Flume Commands
Refer to the file Module-9 Demo 3 provided in the LMS for all the steps in detail
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
93 / 118
Demo 4 – Setting up an Agent
Refer to the file Module-9 Demo 4 provided in the LMS for all the steps in detail
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023

Presentation 94 / 118

Setting up an Agent

- Flume agents are configured using plain text configuration file.
- Flume configuration uses the Java properties file format, which is simply a plain text file with new line separated key-value pairs.
- Configurations for one or more agents can be specified in the same configuration file.
- The configuration file includes properties of each source, sink and channel in an agent and how they are wired together to form data flows.

Open flume-conf.properties
/usr/lib/apache-flume-1.7.0-bin/conf

```
# The configuration file needs to define the sources,
# the channels and the sinks.
# Sources, channels and sinks are defined per agent,
# in this case called 'agent'

agent.sources = seqGenSrc
agent.channels = memoryChannel
agent.sinks = loggerSink

# For each one of the sources, the type is defined
agent.sources.seqGenSrc.type = seq

# The channel can be defined as follows.
agent.sources.seqGenSrc.channels = memoryChannel

# Each sink's type must be defined
agent.sinks.loggerSink.type = logger

#Specify the channel the sink should use
agent.sinks.loggerSink.channel = memoryChannel

# Each channel's type is defined.
agent.channels.memoryChannel.type = memory

# Other config values specific to each type of channel(sink or source)
# can be defined as well
# In this case, it specifies the capacity of the memory channel
agent.channels.memoryChannel.capacity = 100
```

edureka!

Presentation 95 / 118

Simple Configuration File

Name the components on the agent a1

a1.sources = r1

a1.sinks = k1

a1.channels = c1

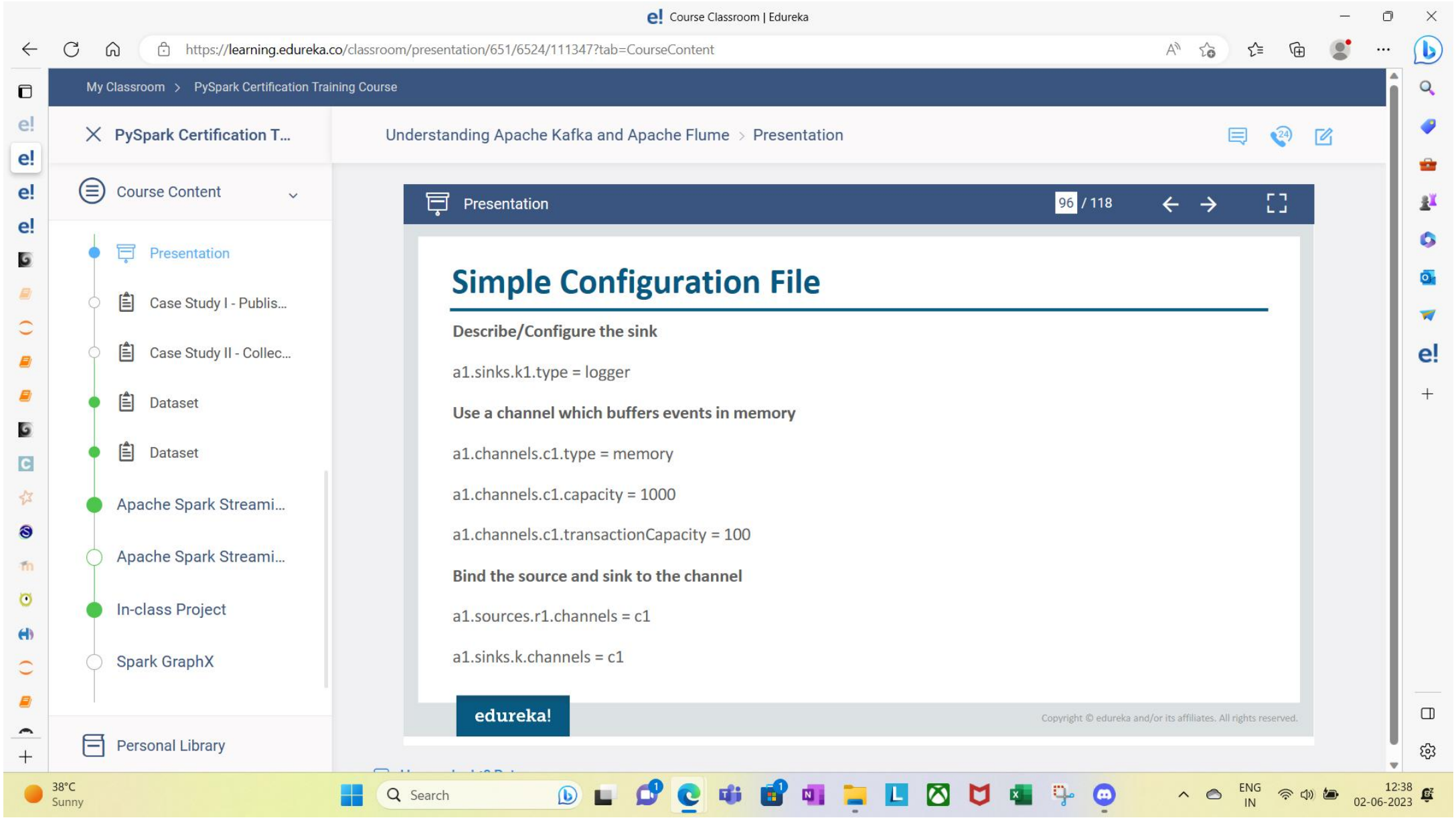
Describe/Configure the source

a1.sources.r1.type = netcat

a1.sources.r1.bind = localhost

a1.sources.r1.port = 44444

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
96 / 118
Simple Configuration File
Describe/Configure the sink
a1.sinks.k1.type = logger
Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100
Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k.channels = c1
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023

Presentation 97 / 118

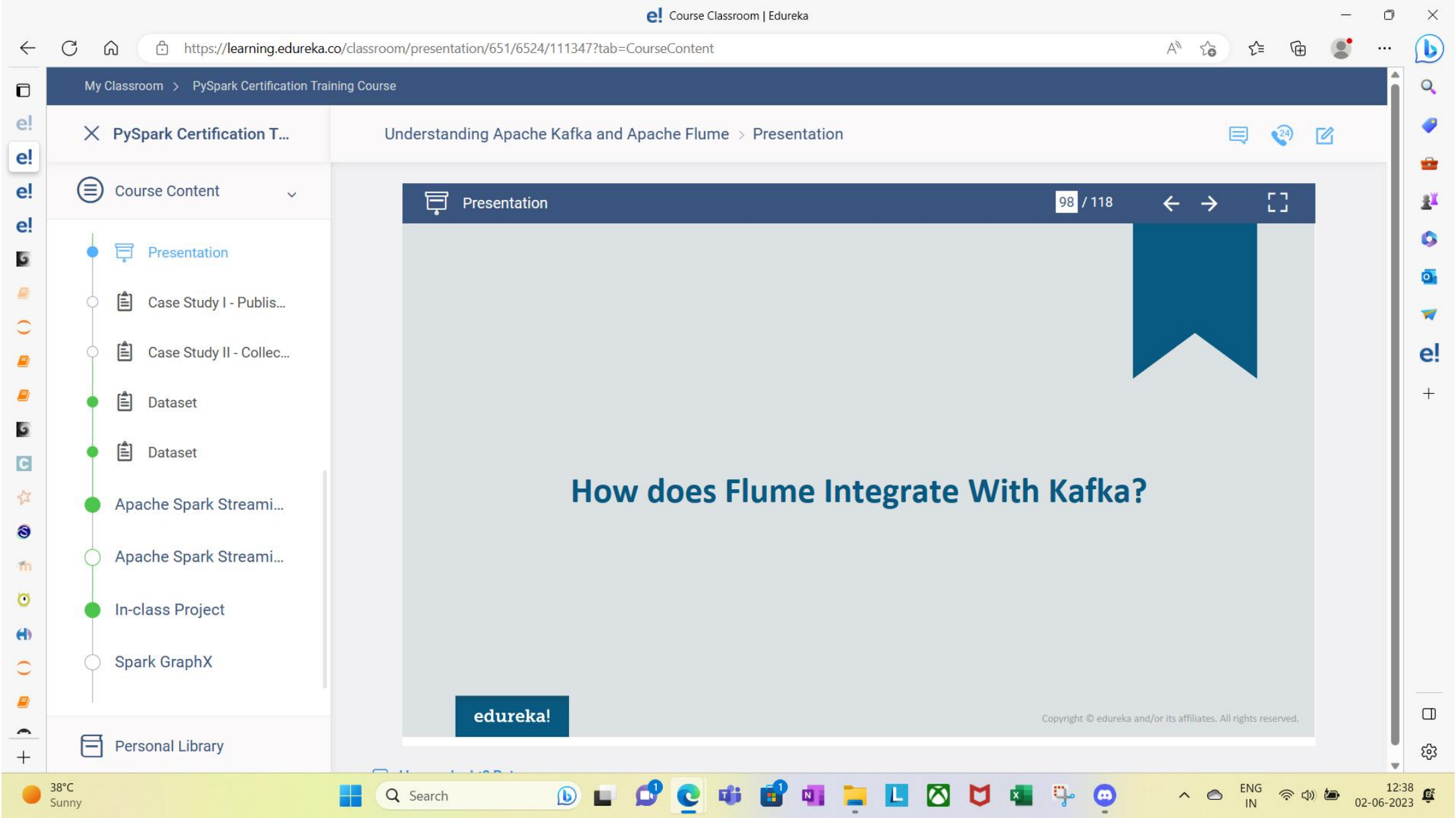
Starting a Flume Agent

- An agent is started using a shell script called flume-ng which is located in the bin directory of the Flume distribution.
- We need to specify the agent name, the config directory, and the config file on the command line:

```
$ bin/flume-ng agent --conf conf --conf-file flume.conf --name a1
-Dflume.root.logger=INFO,console
```

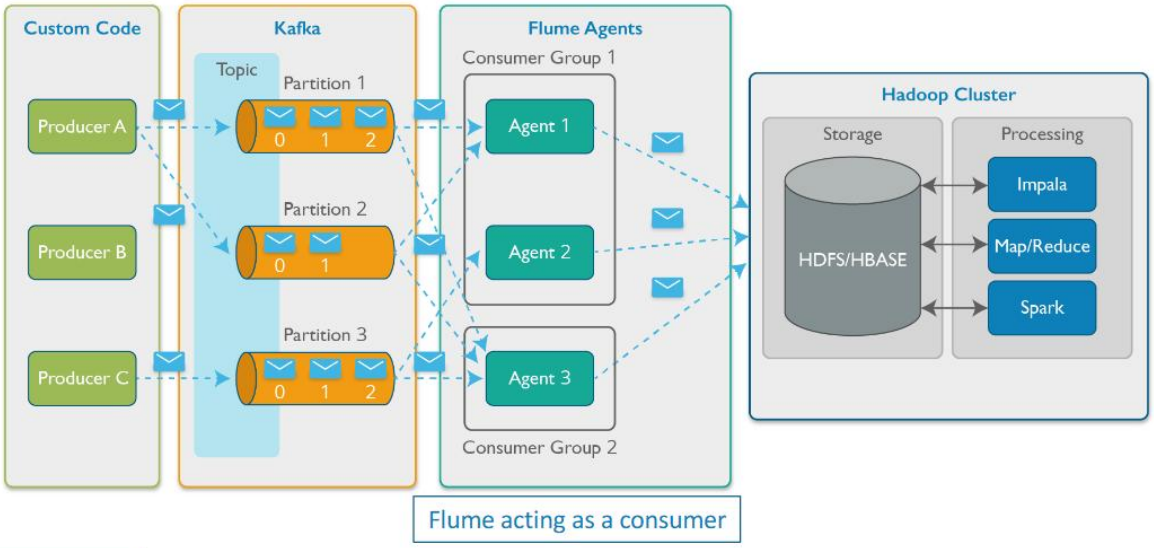
Parameter	Description
agent	Command to start the Flume agent
--conf , -c<conf>	Use configuration file in the conf directory
-f<file>	Specifies a config file path
--name, -n<name>	Name of the agent
-D property=value	Sets a Java system property value

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
98 / 118
How does Flume Integrate With Kafka?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023

Presentation 99 / 118

How Flume Integrates With Kafka?

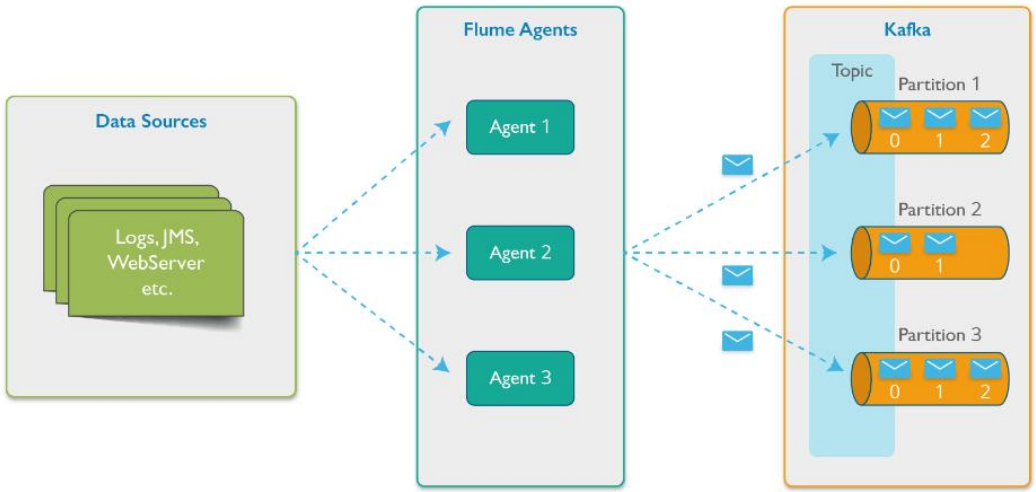


Flume acting as a consumer

edureka!

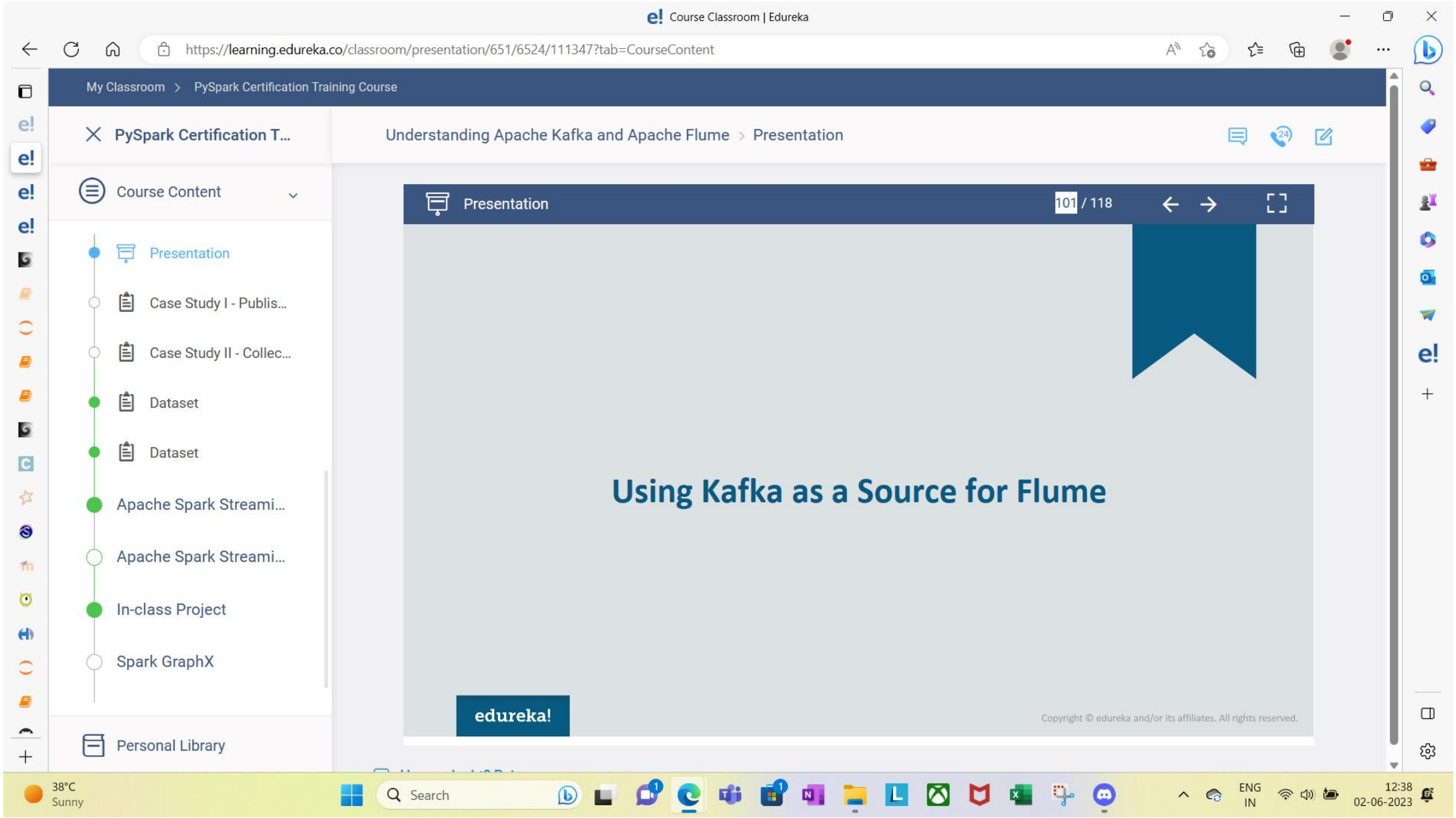
Presentation 100 / 118

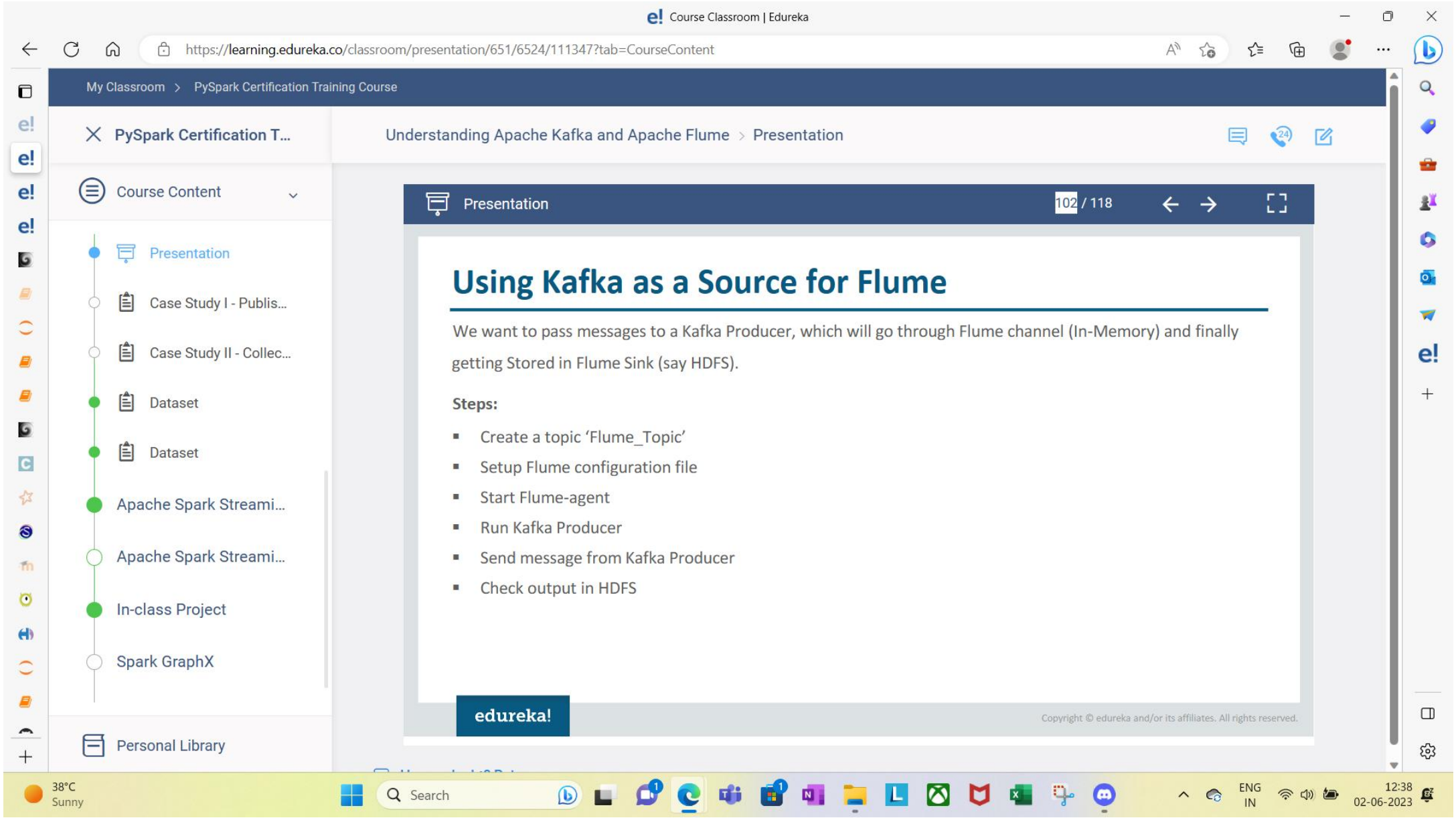
How Flume Integrates With Kafka?



Flume acting as a producer

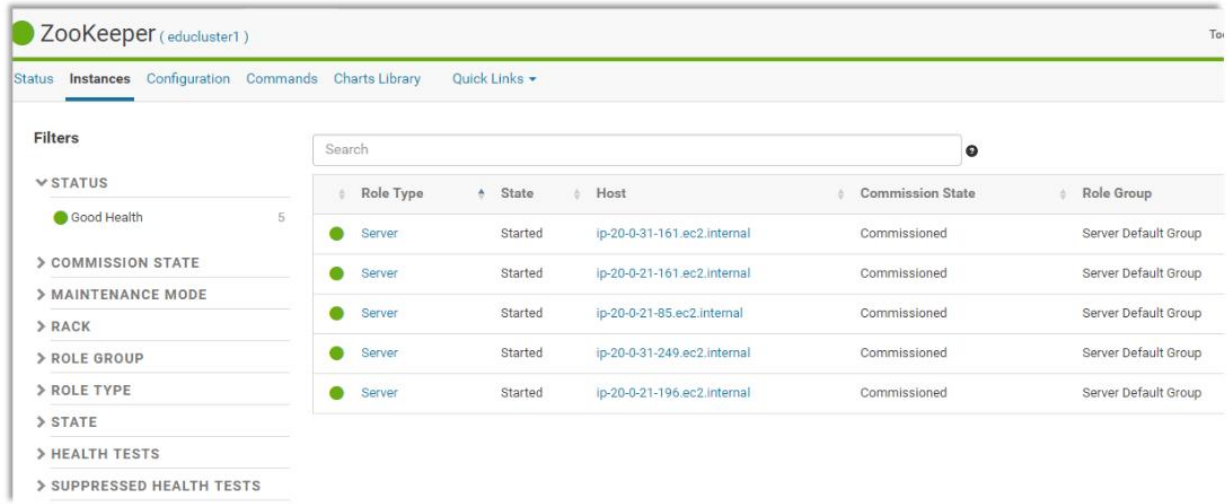
edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
101 / 118
Using Kafka as a Source for Flume
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
102 / 118
Using Kafka as a Source for Flume
We want to pass messages to a Kafka Producer, which will go through Flume channel (In-Memory) and finally getting Stored in Flume Sink (say HDFS).
Steps:
Create a topic 'Flume_Topic'
Setup Flume configuration file
Start Flume-agent
Run Kafka Producer
Send message from Kafka Producer
Check output in HDFS
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023

Presentation 103 / 118

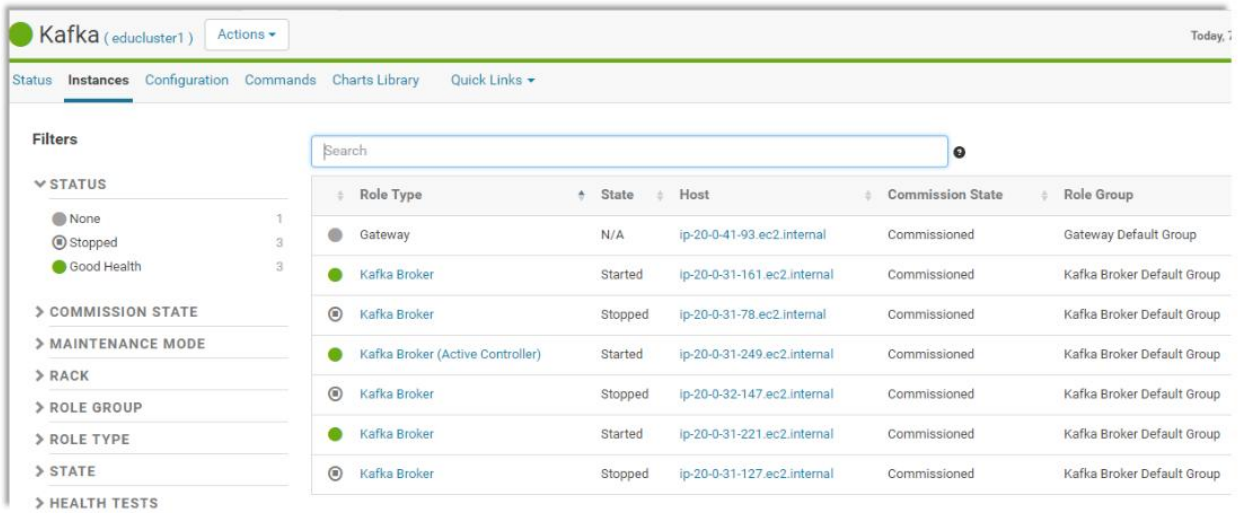
Monitoring Zookeeper Server



edureka!

Presentation 104 / 118

Monitoring Kafka Server



Presentation 105 / 118

Create Kafka Topic

Command: `kafka-topics --create --zookeeper ip-20-0-21-161.ec2.internal:2181 --replication-factor 1 --partitions 1 --topic Flume_Topic`

```
[edureka_249489@ip-20-0-41-190 ~]$ kafka-topics --create --zookeeper ip-20-0-21-161.ec2.internal:2181 --replication-factor 1 --partitions 1 --topic Flume_Topic

18/02/16 12:42:35 INFO zkclient.ZkClient: zookeeper state changed (SyncConnected)
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could collide. To avoid issues it is best to use either, but not both.
18/02/16 12:42:36 INFO admin.AdminUtils$: Topic creation {"version":1,"partitions":{"0":[298]}}
Created topic "Flume_Topic".
18/02/16 12:42:36 INFO zkclient.ZkEventThread: Terminate ZkClient event thread.
18/02/16 12:42:36 INFO zookeeper.ZooKeeper: Session: 0x76166221525602b closed
18/02/16 12:42:36 INFO zookeeper.ClientCnxn: EventThread shut down for session: 0x76166221525602b
```

edureka!

Presentation 106 / 118

Setup Flume Configuration File

flumexercise.conf - Notepad
File Edit Format View Help

```
agent1.sources  = source1
agent1.channels = channel1
agent1.sinks = sink1

agent1.sources.source1.type = org.apache.flume.source.kafka.KafkaSource
agent1.sources.source1.kafka.bootstrap.servers = ip-20-0-31-210.ec2.internal:9092
agent1.sources.source1.kafka.topics = Flume_Topic
agent1.sources.source1.kafka.consumer.group.id = flume
agent1.sources.source1.channels = channel1
agent1.sources.source1.interceptors = i1
agent1.sources.source1.interceptors.i1.type = timestamp
agent1.sources.source1.kafka.consumer.timeout.ms = 100

agent1.channels.channel1.type = memory
agent1.channels.channel1.capacity = 10000
agent1.channels.channel1.transactionCapacity = 1000

agent1.sinks.sink1.type = hdfs
agent1.sinks.sink1.hdfs.path = hdfs://nameservice1/user/edureka_249489/Flume_Kafka
agent1.sinks.sink1.hdfs.rollInterval = 5
agent1.sinks.sink1.hdfs.rollSize = 0
agent1.sinks.sink1.hdfs.rollCount = 0
agent1.sinks.sink1.hdfs.fileType = DataStream
agent1.sinks.sink1.channel = channel1
```

edureka!

Presentation 107 / 118

Start Flume Agent

Command: `flume-ng agent --conf conf --conf-file flumexercise.conf --name agent1 -Dflume.root.logger=INFO,console`

```
[edureka_249489@ip-20-0-41-190 ~]flume-ng agent --conf conf --conf-file flumexercise.conf --name agent1 -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
ib/*:/etc/hadoop/conf:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop/lib/*:/opt/cloudera/parcels/CDH-5.11.
/hadoop/libexec/../../hadoop/.//*:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop-hdfs/./:/opt/cloudera/par
.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/
cels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libexec/../../hadoop-yarn/lib/*:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/libe
.//*:/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/lib/*:/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/.//*:/opt/cloudera/parcels/CDH-5.11.
/hbase/bin/../conf:/usr/java/jdk1.8.0_144-cloudera/lib/tools.jar:/opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hbase/bin/..:/opt/clou
```

edureka!

Presentation 108 / 118

Run Kafka Producer

Command: `kafka-console-producer --broker-list ip-20-0-31-210.ec2.internal:9092 --topic Flume_Topic`

```
[edureka_249489@ip-20-0-41-190 ~]$ kafka-console-producer --broker-list ip-20-0-31-210.ec2.internal:9092 --topic Flume_Topic
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/s
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/sl
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/02/16 12:57:38 INFO producer.ProducerConfig: ProducerConfig values:
```

edureka!

Presentation 109 / 118

Send Message From Kafka Producer

```
18/02/16 12:57:38 INFO utils.AppInfoParser: Kafka version : 0.11.0-kafka-3.0.0
18/02/16 12:57:38 INFO utils.AppInfoParser: Kafka commitId : unknown
>This is a test for kafka and flume integration
>Welcome to Flume
>
```

edureka!

Presentation 110 / 118

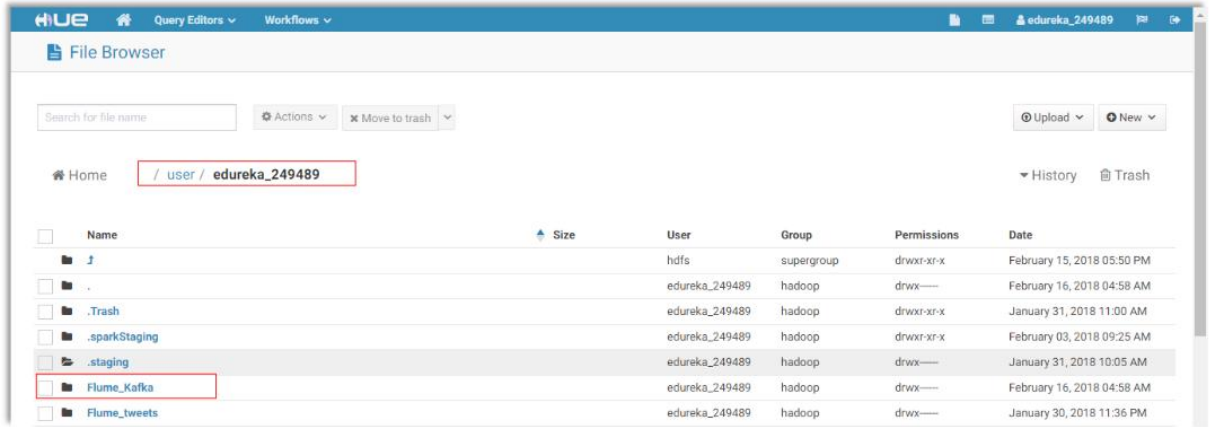
Check Flume Agent window

```
18/02/16 12:55:32 WARN consumer.ConsumerConfig: The configuration timeout.ms = 100 was supplied but isn't a known config.
18/02/16 12:55:32 INFO utils.AppInfoParser: Kafka version : 0.9.0-kafka-2.0.2
18/02/16 12:55:32 INFO utils.AppInfoParser: Kafka commitId : unknown
18/02/16 12:55:32 INFO internals.AbstractCoordinator: Discovered coordinator ip-20-0-31-221.ec2.internal:9092 (id: 2147483350) for group flume.
18/02/16 12:55:32 INFO internals.ConsumerCoordinator: Revoking previously assigned partitions [] for group flume
18/02/16 12:55:32 INFO internals.AbstractCoordinator: (Re-)joining group flume
18/02/16 12:55:35 INFO internals.AbstractCoordinator: Successfully joined group flume with generation 5
18/02/16 12:55:35 INFO internals.ConsumerCoordinator: Setting newly assigned partitions [Flume_Topic-0] for group flume
18/02/16 12:55:35 INFO kafka.SourceRebalanceListener: topic Flume_Topic - partition 0 assigned.
18/02/16 12:55:35 INFO kafka.KafkaSource: Kafka source source1 started.
18/02/16 12:55:35 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: source1: Successfully registered new MBean.
18/02/16 12:55:35 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: source1 started
18/02/16 12:58:50 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/02/16 12:58:51 INFO hdfs.BucketWriter: Creating hdfs://nameservice1/user/edureka_249489/Flume_Kafka/FlumeData.1518785930893.tmp
18/02/16 12:58:56 INFO hdfs.BucketWriter: Closing hdfs://nameservice1/user/edureka_249489/Flume_Kafka/FlumeData.1518785930893.tmp
18/02/16 12:58:56 INFO hdfs.BucketWriter: Renaming hdfs://nameservice1/user/edureka_249489/Flume_Kafka/FlumeData.1518785930893.tmp to hdfs://nameservice1/user/edureka_249489/Flume_Kafka/FlumeData.1518785930893
18/02/16 12:58:56 INFO hdfs.HDFSEventSink: Writer callback called.
```

edureka!

Presentation 111 / 118

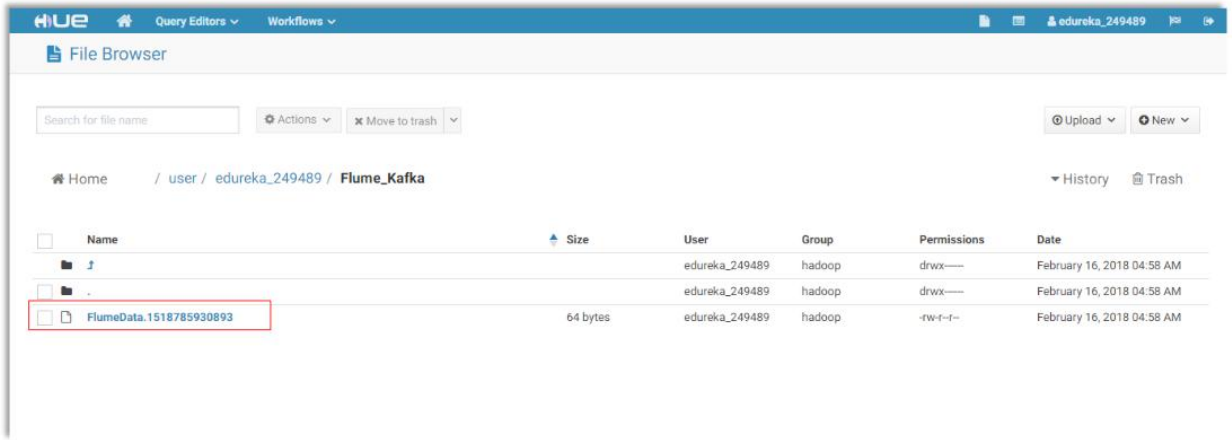
Browse to the HDFS Directory



edureka!

Presentation 112 / 118

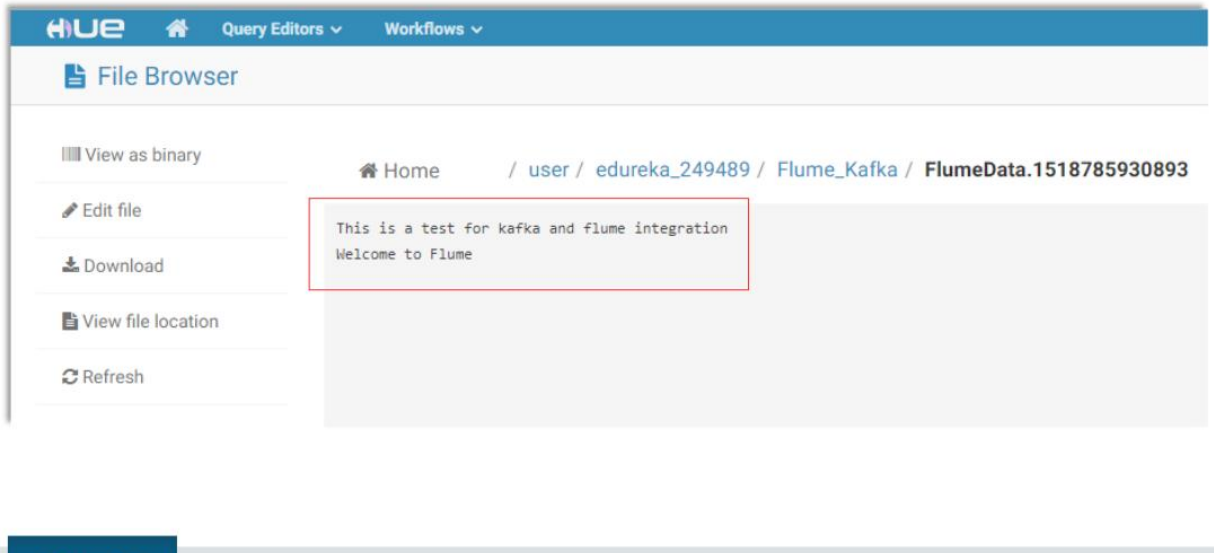
Open the File Created



edureka!

Presentation 113 / 118

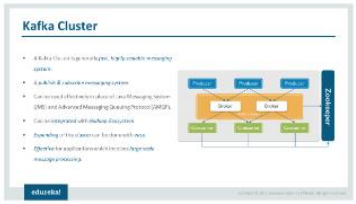
Check Output

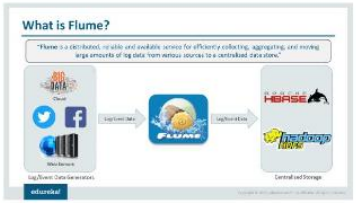


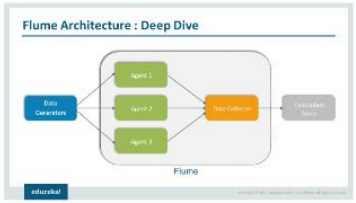
edureka!

Presentation 114 / 118

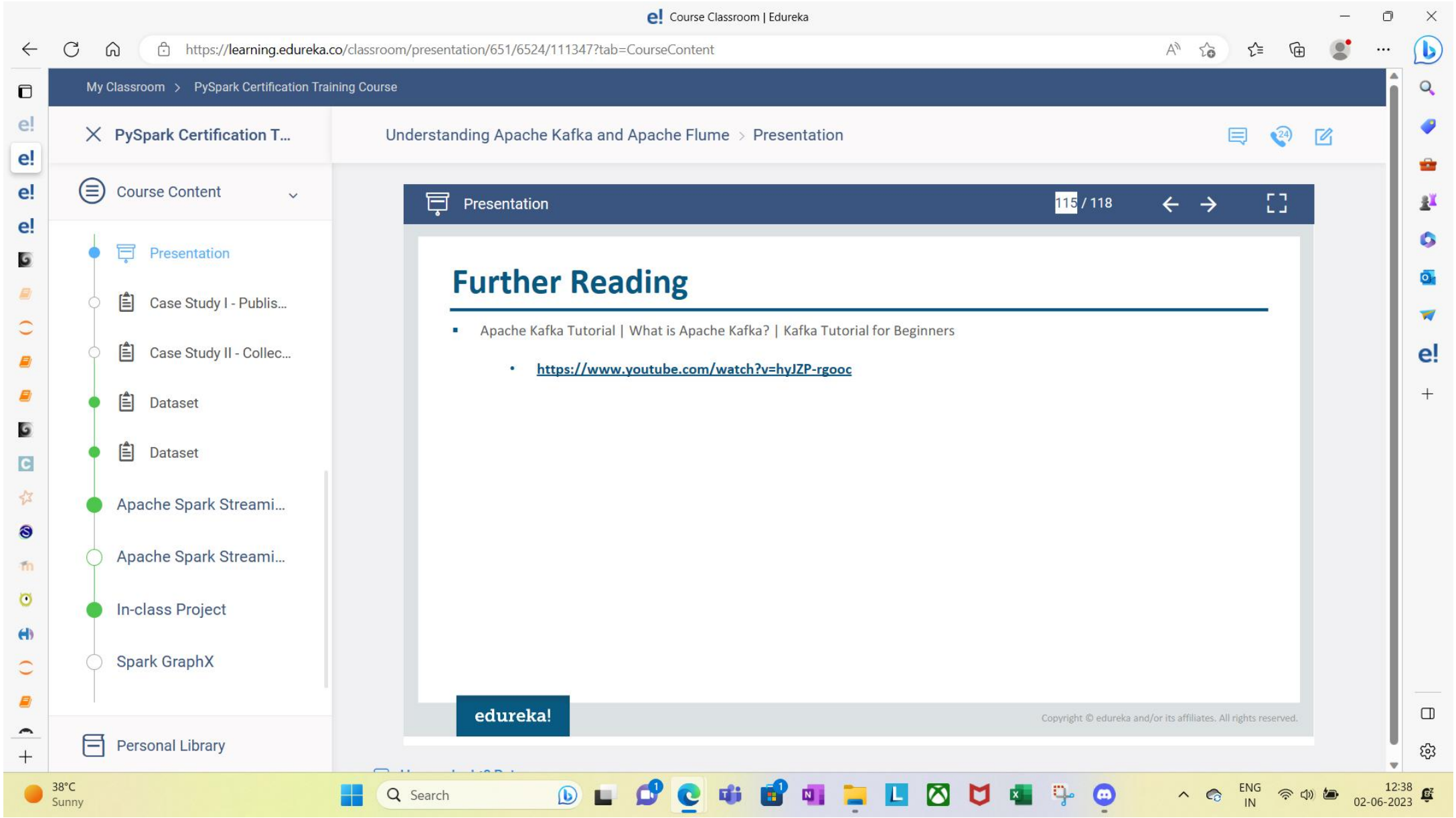
Summary







edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6524/111347?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Understanding Apache Kafka and Apache Flume > Presentation
Course Content
Presentation
Case Study I - Publis...
Case Study II - Collec...
Dataset
Dataset
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Personal Library
Presentation
115 / 118
Further Reading
Apache Kafka Tutorial | What is Apache Kafka? | Kafka Tutorial for Beginners
https://www.youtube.com/watch?v=hyJZP-rgooc
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
38°C
Sunny
Search
ENG
IN
12:38
02-06-2023