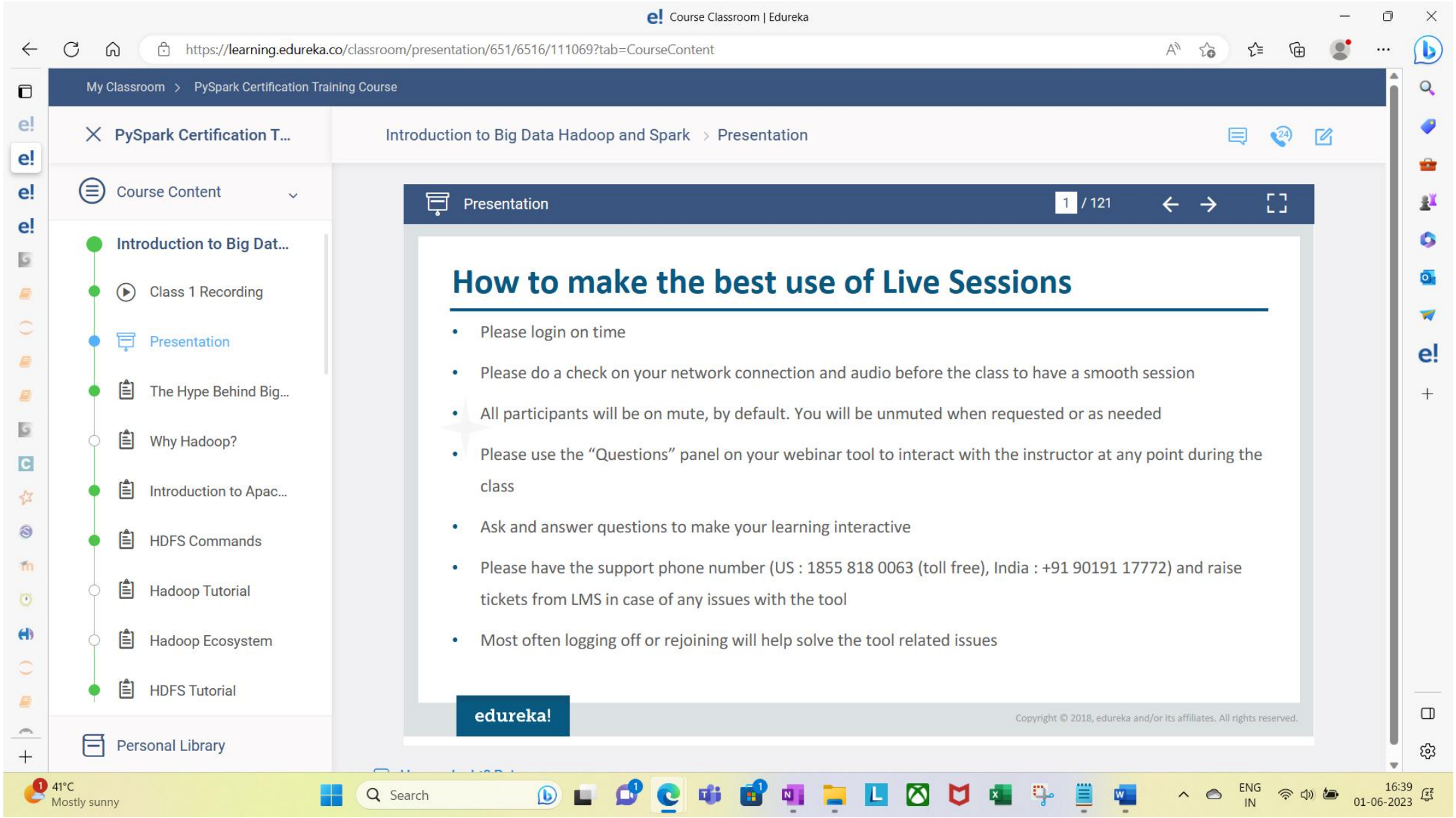
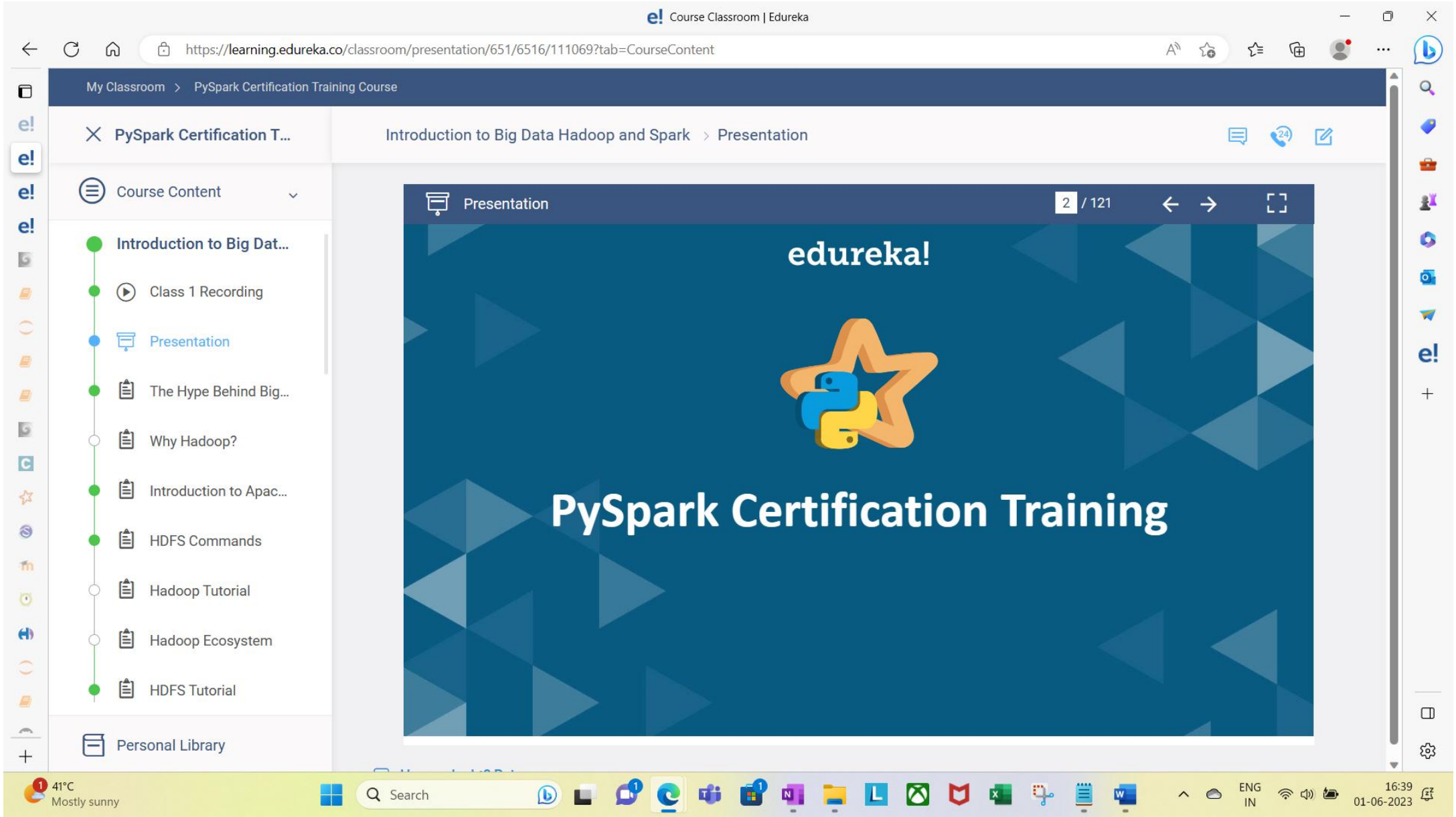

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
1 / 121
How to make the best use of Live Sessions
Please login on time
Please do a check on your network connection and audio before the class to have a smooth session
All participants will be on mute, by default. You will be unmuted when requested or as needed
Please use the "Questions" panel on your webinar tool to interact with the instructor at any point during the class
Ask and answer questions to make your learning interactive
Please have the support phone number (US : 1855 818 0063 (toll free), India : +91 90191 17772) and raise tickets from LMS in case of any issues with the tool
Most often logging off or rejoining will help solve the tool related issues
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Mostly sunny
Search
ENG
IN
16:39
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
2 / 121
edureka!
PySpark Certification Training
41°C
Mostly sunny
Search
ENG
IN
16:39
01-06-2023

Presentation 3 / 121

COURSE OUTLINE

MODULE 01

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

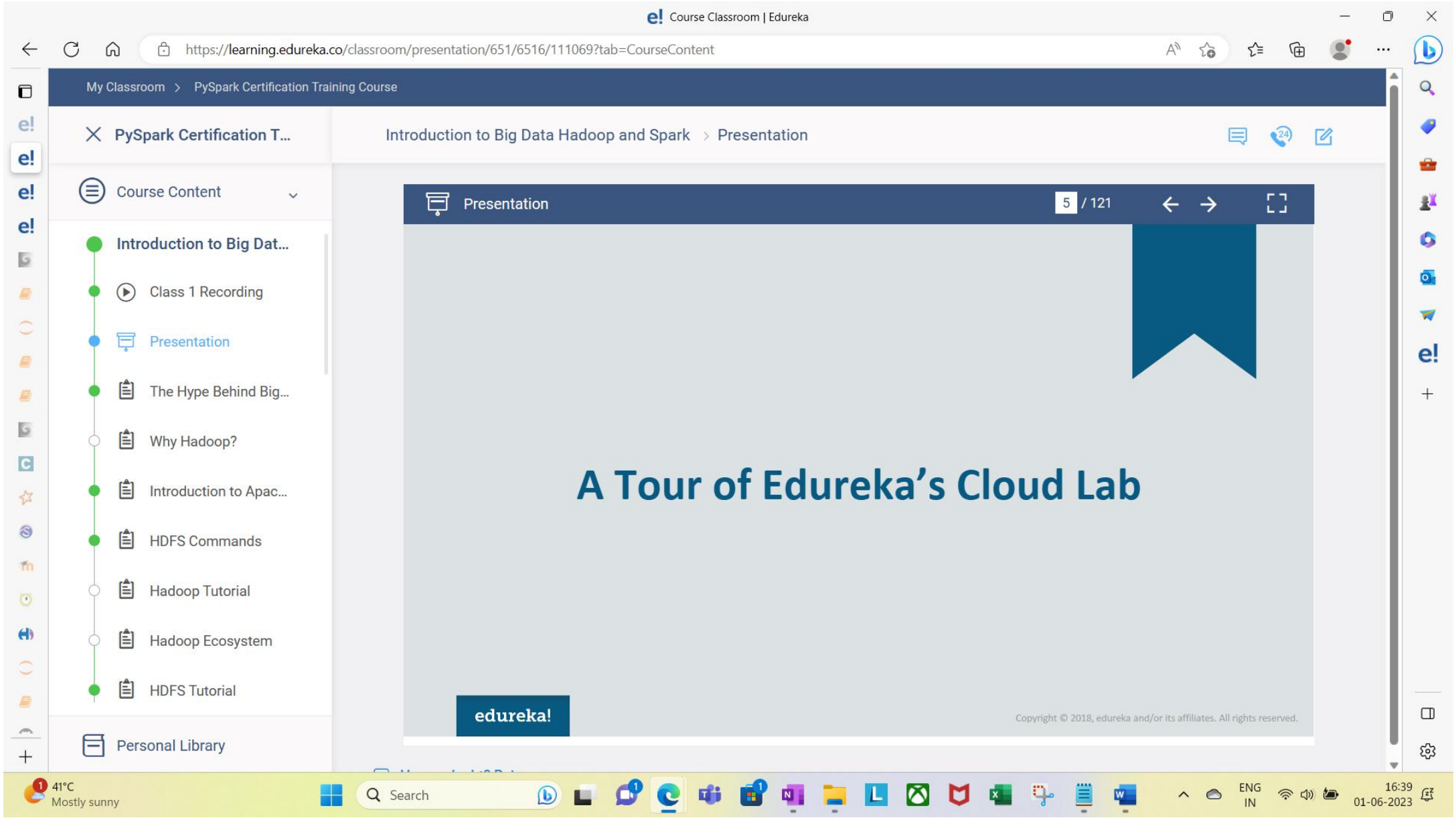
Presentation 4 / 121

Objectives

After completing this module, you should be able to:

- Evaluate present Structure of Data and problems
- Describe Big Data and it's domain
- Explain Hadoop Core Components and HDFS
- Understand Replication and Rack Awareness
- Explain YARN and it's components
- Differentiate between Big Data Analytics with Batch and Real-Time Processing
- Understand the need for Spark and explain what is Spark
- Analyse how Spark differs from its competitors
- Recognize Spark's place in Hadoop ecosystem

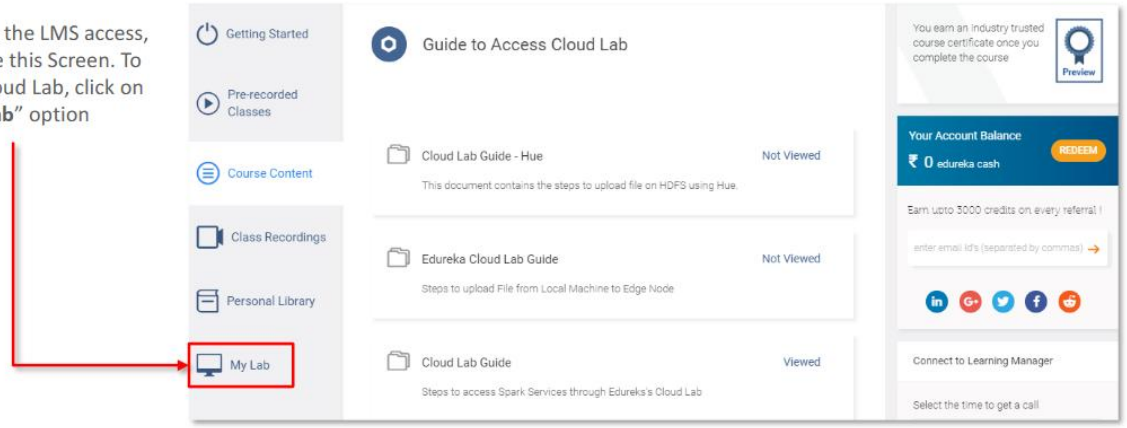
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
5 / 121
A Tour of Edureka's Cloud Lab
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Mostly sunny
Search
ENG
IN
16:39
01-06-2023

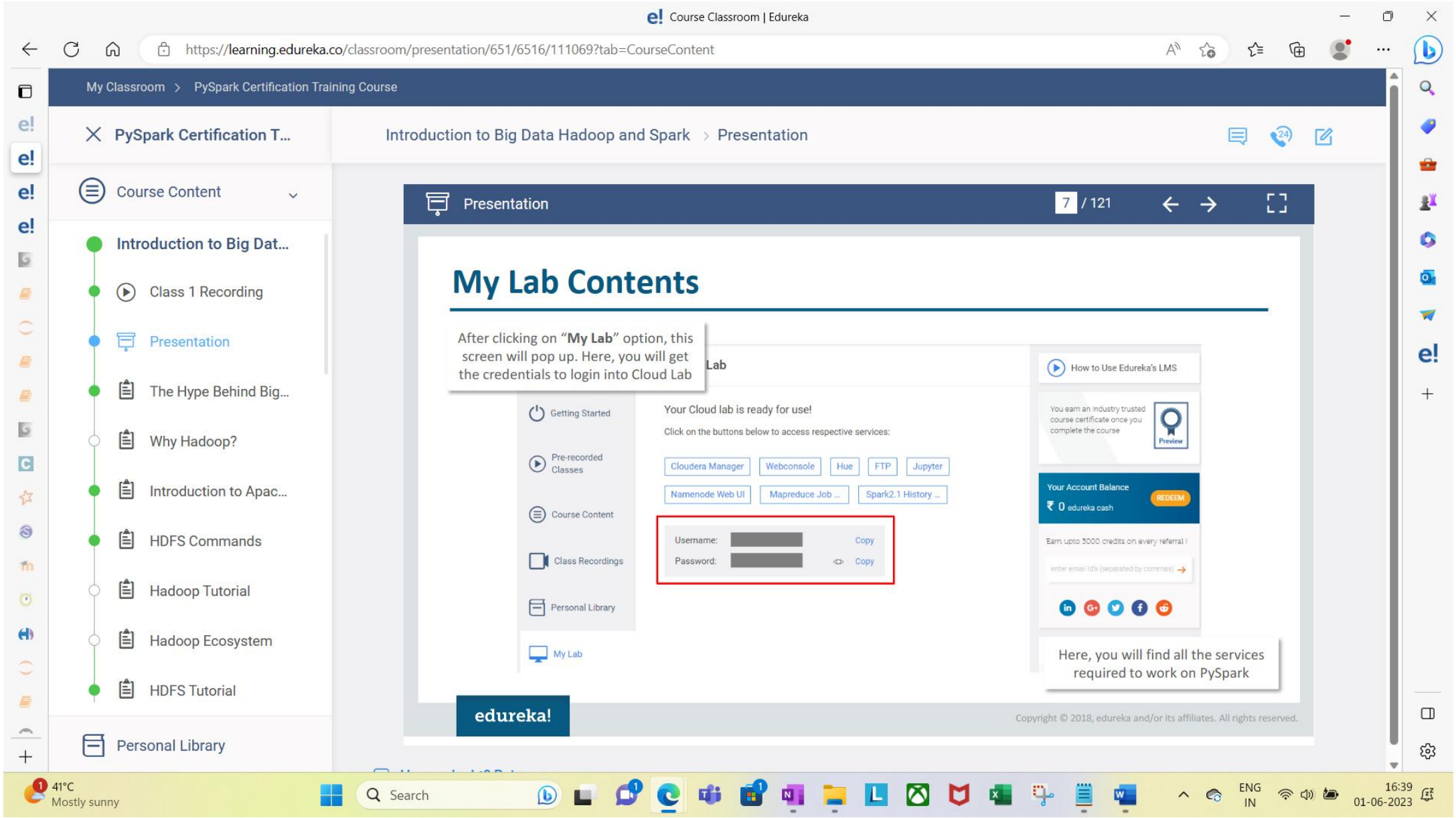
Presentation 6 / 121

LMS Overview

Once you get the LMS access, you will have this Screen. To login into Cloud Lab, click on "**My Lab**" option



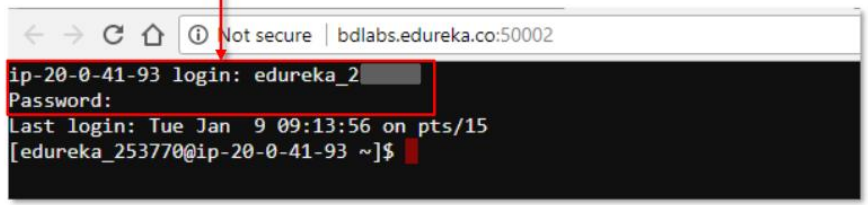
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
7 / 121
My Lab Contents
After clicking on "My Lab" option, this screen will pop up. Here, you will get the credentials to login into Cloud Lab
Lab
Getting Started
Pre-recorded Classes
Course Content
Class Recordings
Personal Library
My Lab
Your Cloud lab is ready for use!
Click on the buttons below to access respective services:
Cloudera Manager
Webconsole
Hue
FTP
Jupyter
Namenode Web UI
Mapreduce Job ...
Spark2.1 History ...
Username:
Password:
Copy
Copy
How to Use Edureka's LMS
You earn an industry trusted course certificate once you complete the course
Preview
Your Account Balance
₹ 0 edureka cash
REDEEM
Earn upto 3000 credits on every referral !
enter email id's (separated by commas)
Here, you will find all the services required to work on PySpark
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Mostly sunny
Search
ENG
IN
16:39
01-06-2023

Presentation 8 / 121

WebConsole

WebConsole is terminal. You have to copy-paste the credentials from the previous window to this window, to login into WebConsole

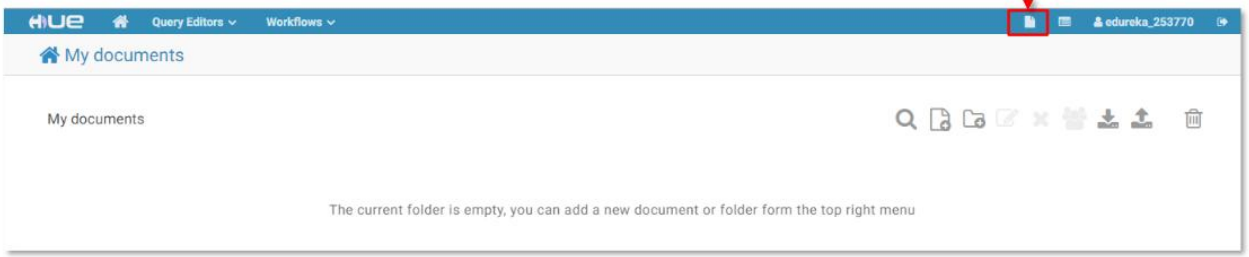


edureka!

Presentation 9 / 121

HUE

To see all the files present in HDFS, click on "**HDFS Browser**"

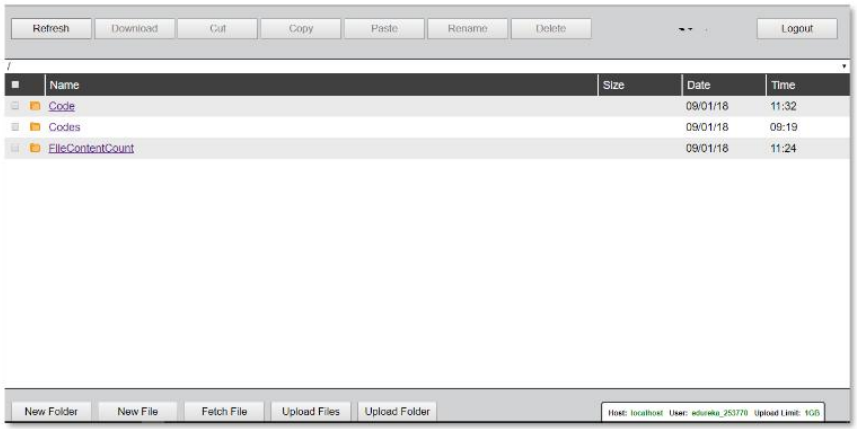


edureka!

Presentation 10 / 121

FTP Server

Using **FTP** you can upload or download any file/ folder from your local system to cloud lab



edureka!

Presentation 11 / 121

Jupyter Notebook

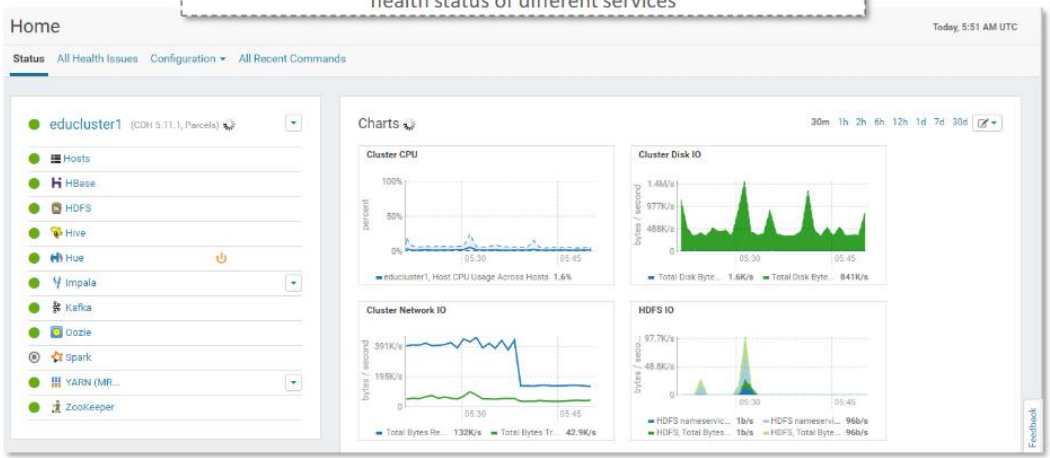
Jupyter Notebook is the IDE used to run all the PySpark Programs

edureka!

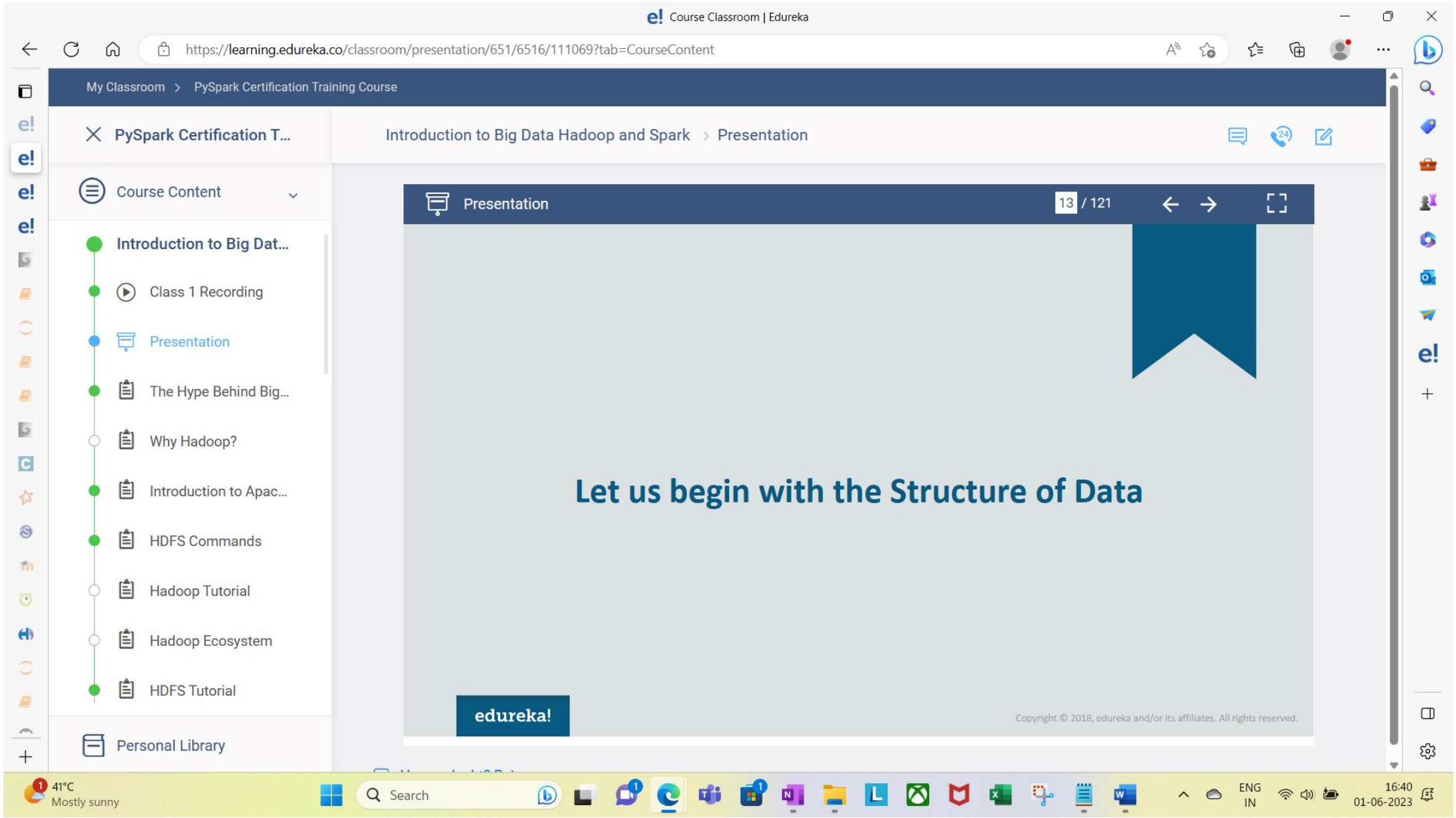
Presentation 12 / 121

Cloudera Manager

Cloudera Manager is used to visualize the configurations and the health status of different services

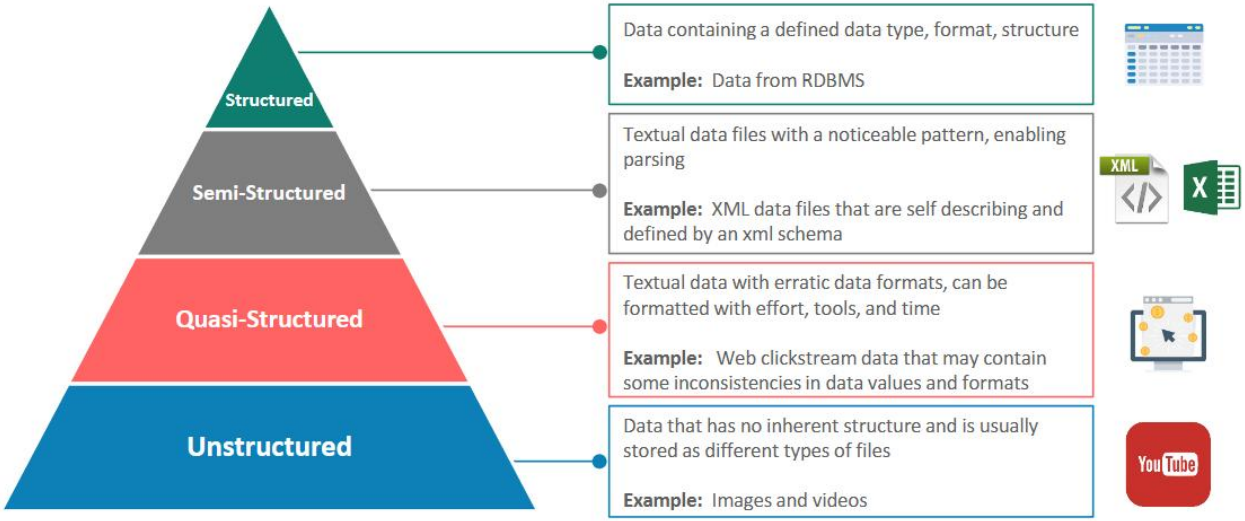


edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
13 / 121
Let us begin with the Structure of Data
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Mostly sunny
Search
ENG
IN
16:40
01-06-2023

Presentation 14 / 121

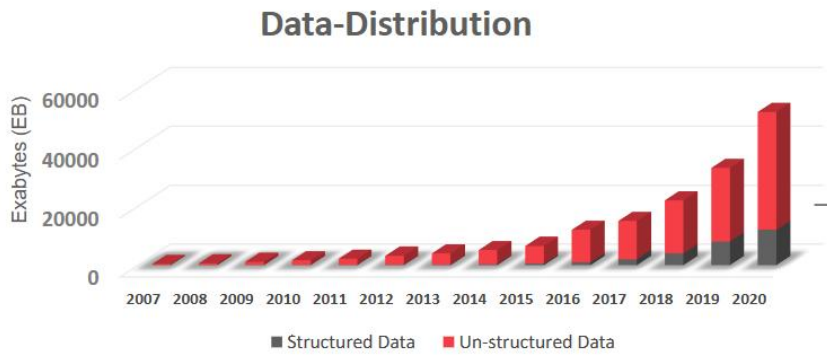
Structure of Data Today



edureka!

Presentation 15 / 121

Unstructured Data is exploding

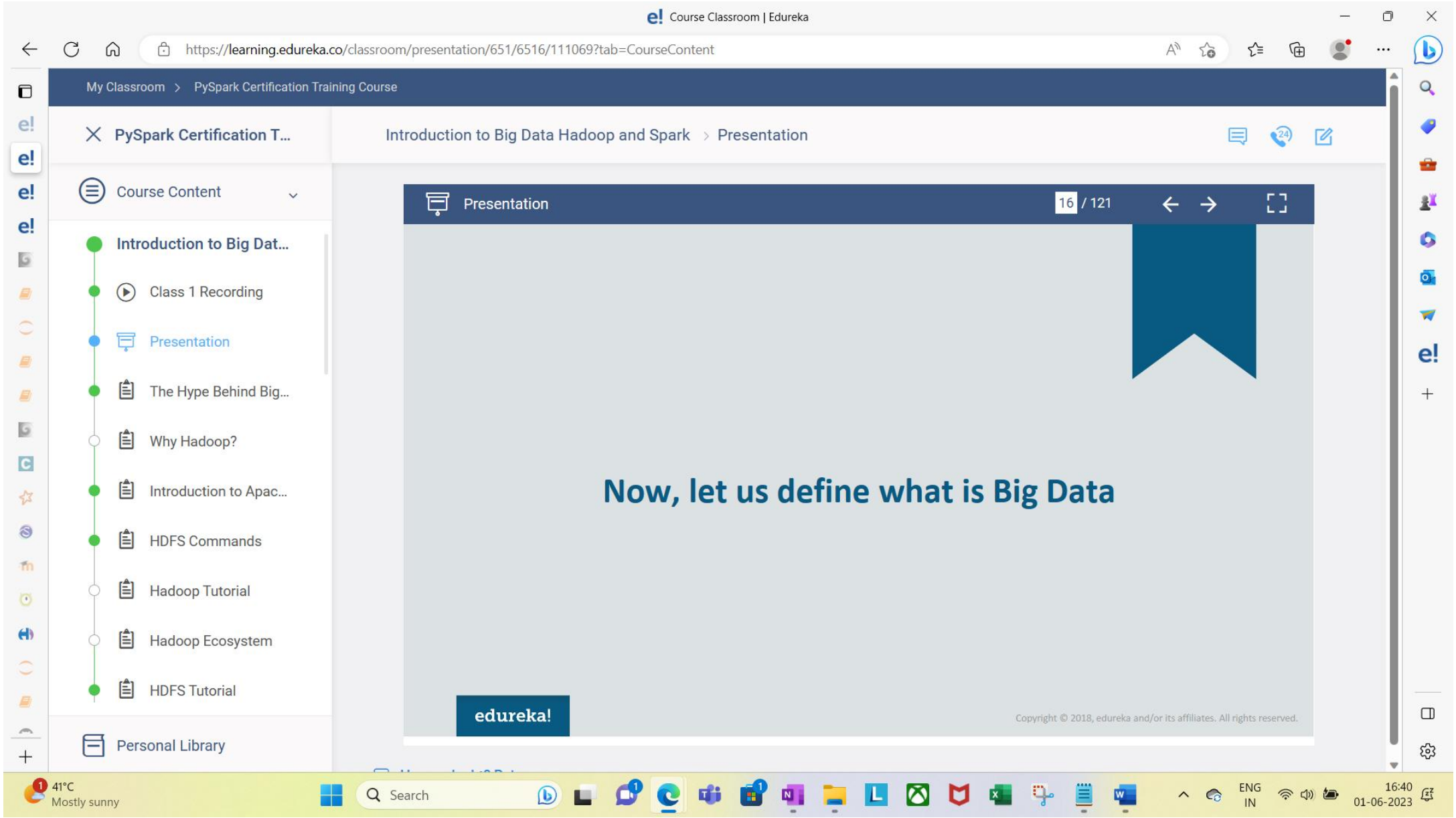


This accelerated growth of **UNSTRUCTURED** data is what turned into **BIG DATA**

By 2020, IDC predicted, Unstructured Data to have reached ***40 Zettabytes (ZB)***

By 2020, there will be ***5,200 GB*** of data for every person on Earth

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
16 / 121
Now, let us define what is Big Data
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Mostly sunny
Search
ENG
IN
16:40
01-06-2023

Presentation 17 / 121

What is Big Data?

BIG DATA is the term for a collection of data sets so ***large and complex*** that it becomes ***difficult*** to process using on-hand database management tools or traditional data processing applications

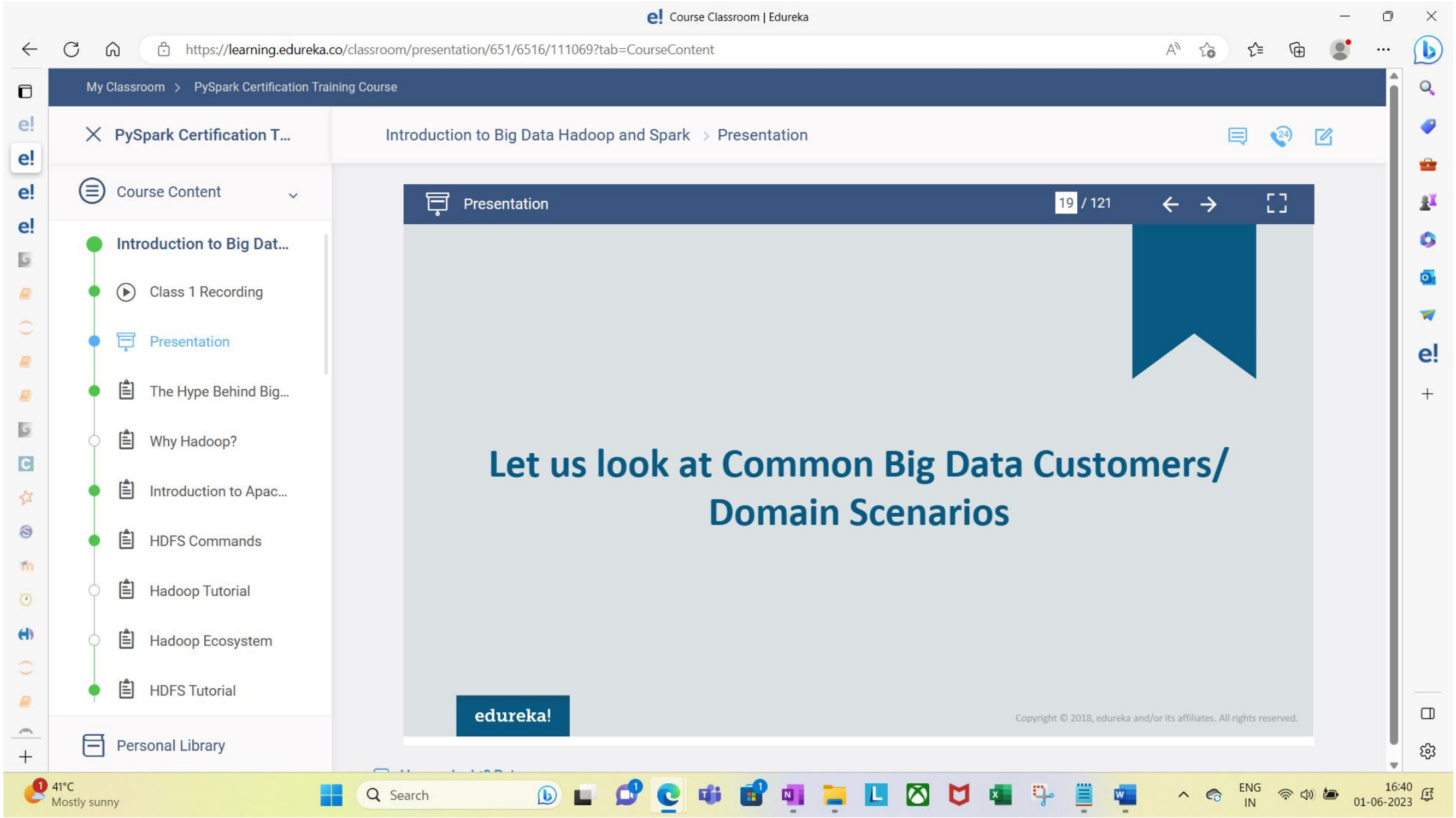
edureka!

Presentation 18 / 121

The 5 V's of Big Data

Volume	Variety	Velocity	Value	Veracity
Processing large amount of data	Processing different types of data	Data is being generated at an alarming rate	Finding correct meaning out of the data	Uncertainty and inconsistencies in the data

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
19 / 121
Let us look at Common Big Data Customers/ Domain Scenarios
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Mostly sunny
Search
ENG
IN
16:40
01-06-2023

Common Big Data Customer/Domain Scenarios

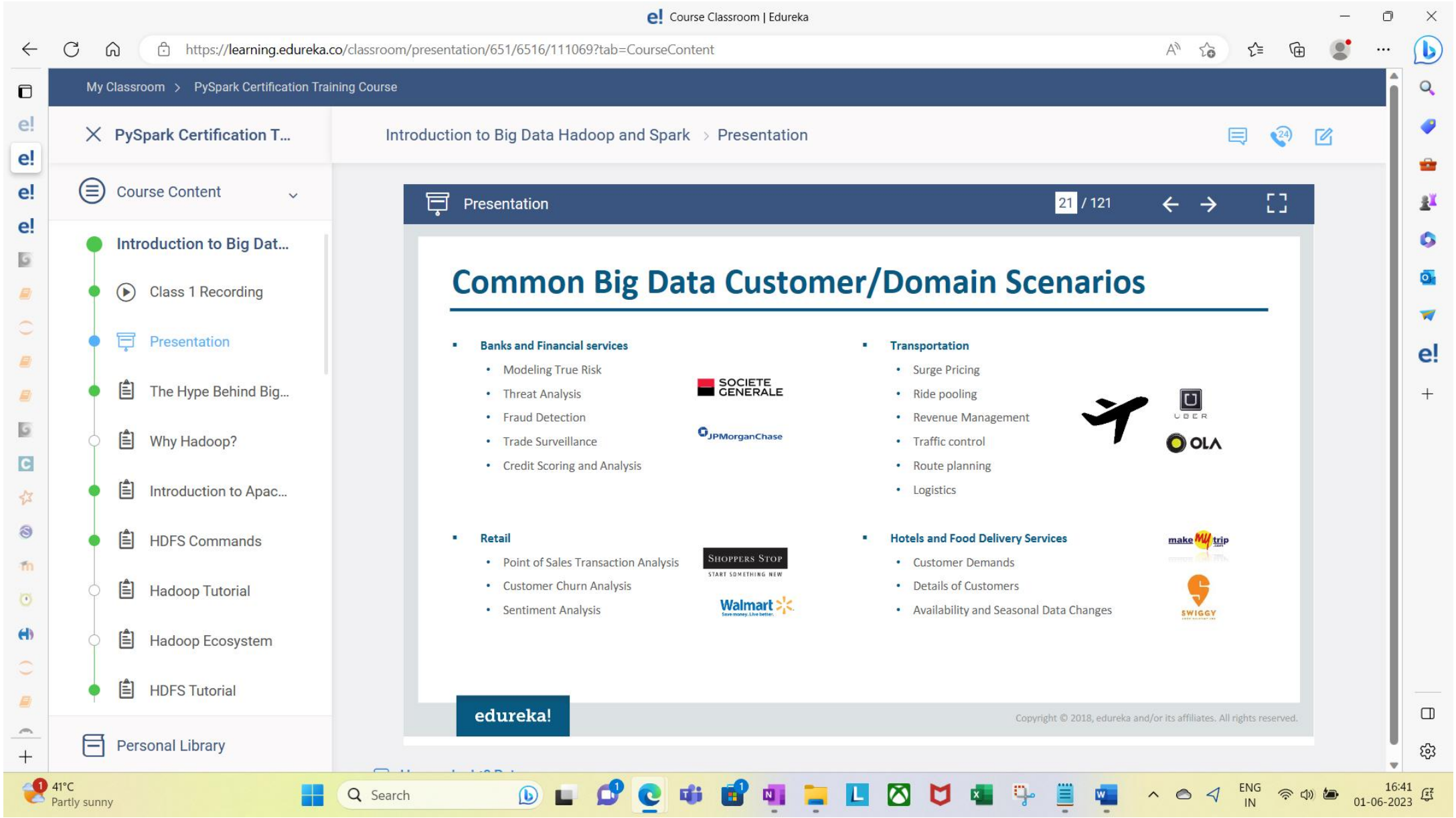
- **Web and e-tailing**
  - Recommendation Engines
  - Ad Targeting
  - Search Quality
  - Abuse and Click Fraud Detection

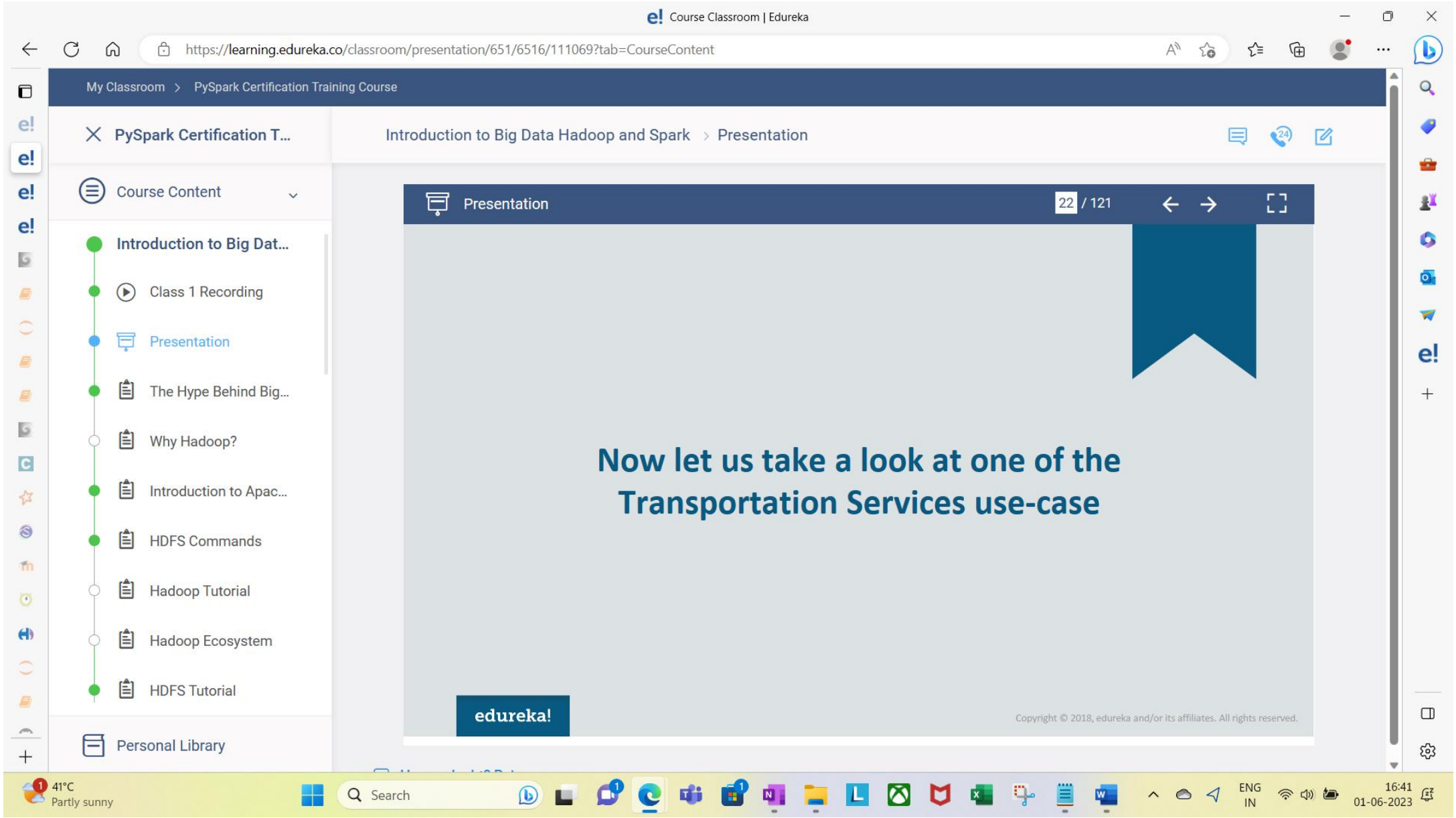
- **Government**
  - Fraud Detection and Cyber Security
  - Welfare Schemes
  - Justice

- **Telecommunications**
  - Customer Churn Prevention
  - Network Performance Optimization
  - Calling Data Record (CDR) Analysis
  - Analysing Network to Predict Failure

- **Healthcare and Life Sciences**
  - Health Information Exchange
  - Gene Sequencing
  - Serialization
  - Healthcare Service Quality Improvements
  - Drug Safety

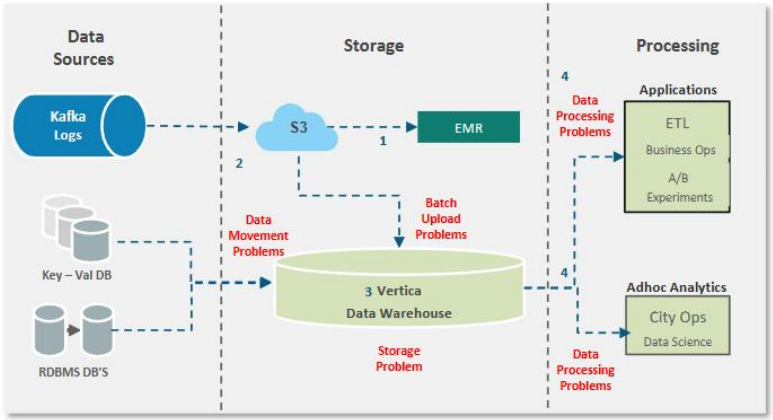
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
21 / 121
Common Big Data Customer/Domain Scenarios
Banks and Financial services
Modeling True Risk
Threat Analysis
Fraud Detection
Trade Surveillance
Credit Scoring and Analysis
SOCIETE GENERALE
JPMorganChase
Transportation
Surge Pricing
Ride pooling
Revenue Management
Traffic control
Route planning
Logistics
UBER
OLA
Retail
Point of Sales Transaction Analysis
Customer Churn Analysis
Sentiment Analysis
SHOPPERS STOP
Walmart
Hotels and Food Delivery Services
Customer Demands
Details of Customers
Availability and Seasonal Data Changes
make MY trip
SWIGGY
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
22 / 121
Now let us take a look at one of the Transportation Services use-case
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 23 / 121

Problems : Old Uber Infrastructure (2014)



PROBLEMS

1. ***Data locality*:** Using a single place for storage led to **data movement problems** and also increased the risk of **Single point of Failure**
2. ***Scale issues*:** As they started growing there were many problems , **mainly with the batch upload problem** as they started collecting many sorts of data like "**Car Speed, traffic data, user data, fare**"
3. ***Storage*:** With the increasing amount of data it became difficult to store the data
4. ***Handling unstructured data*:** With Inconsistency in data (unstructured), the processing problems also increased

edureka!

Probable Solution to the Problems

- We need something which can *store large amounts of data*, and *can be scaled easily* if we get more data

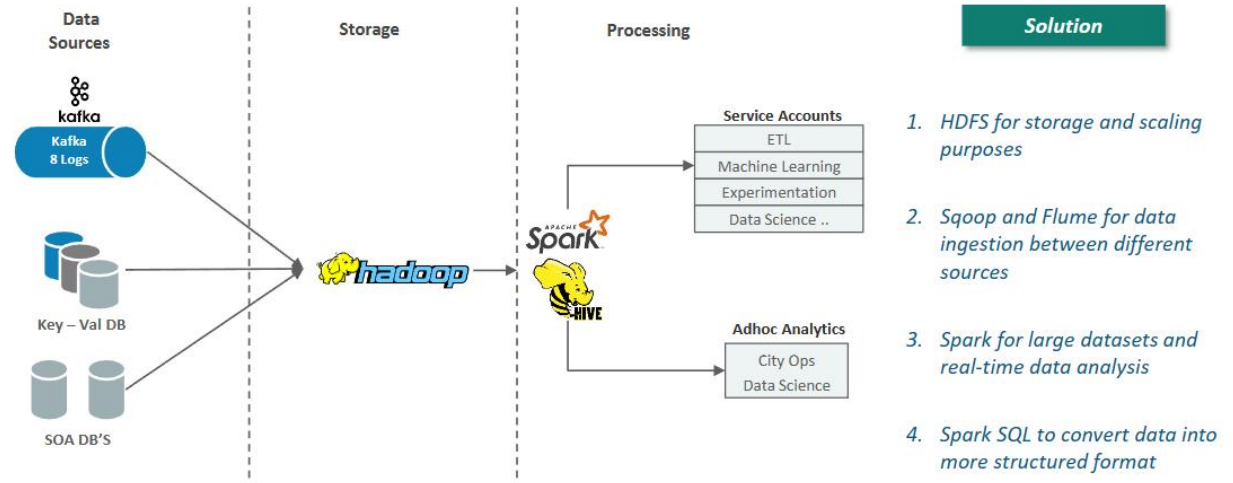
- There shouldn't be any difficulties in *batch-upload of data*
- Data processing and analysis should be fast, *both for static as well as live streaming data*

- Handling and processing unstructured data from different sources into a *structured format*

edureka!

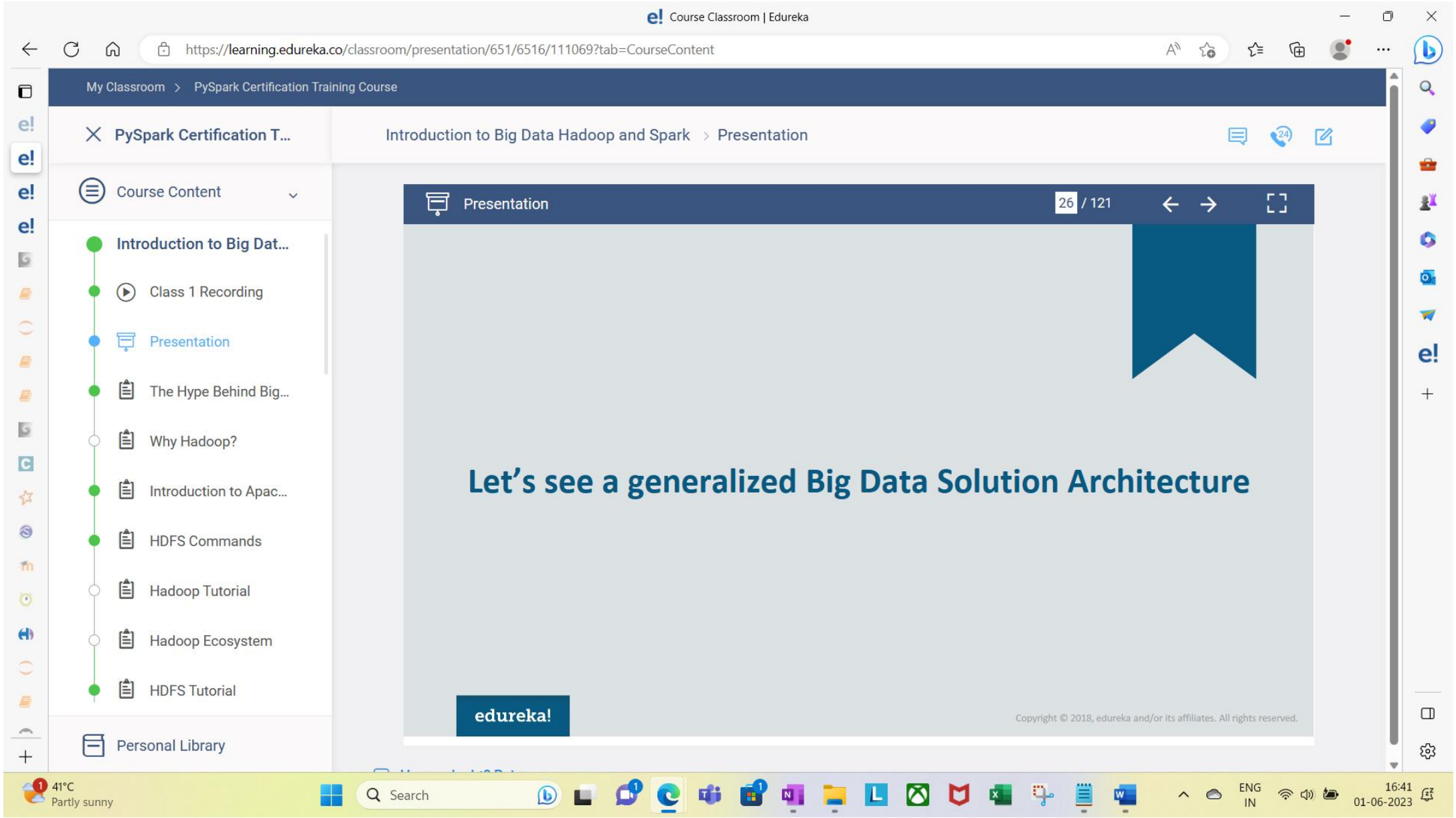
Presentation 25 / 121

Solution : Present Infrastructure



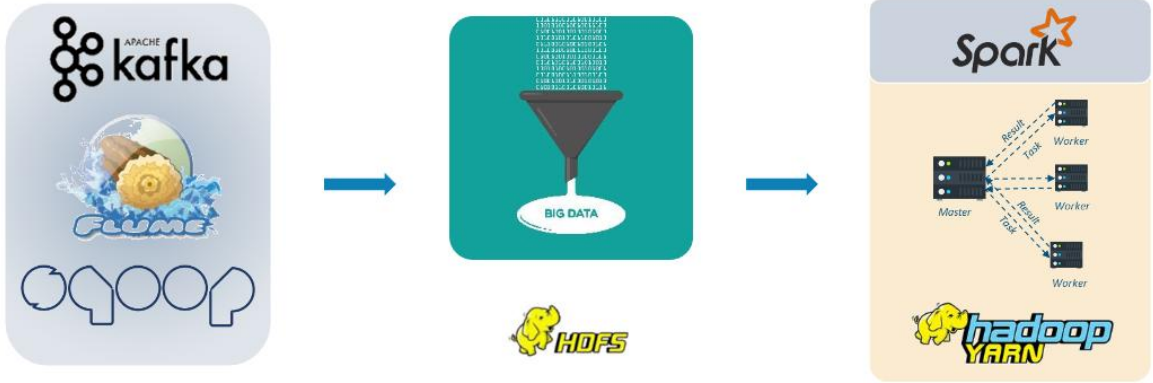
Source : Datanami

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
26 / 121
Let's see a generalized Big Data Solution Architecture
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 27 / 121

Big Data Solution Architecture

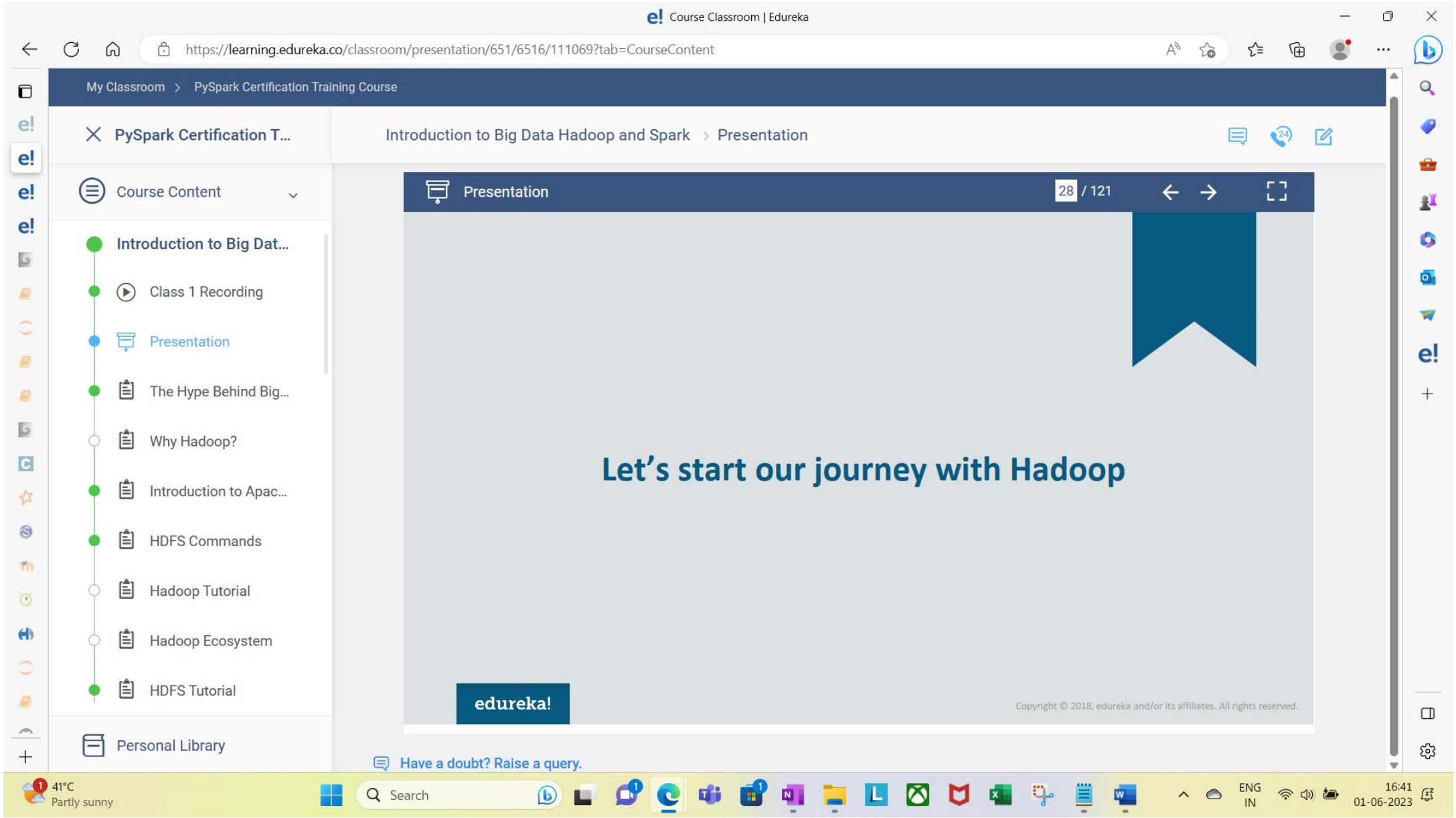


Data sources and ingestion tools

Storing huge & complex data on HDFS

Processing through Hadoop and Spark

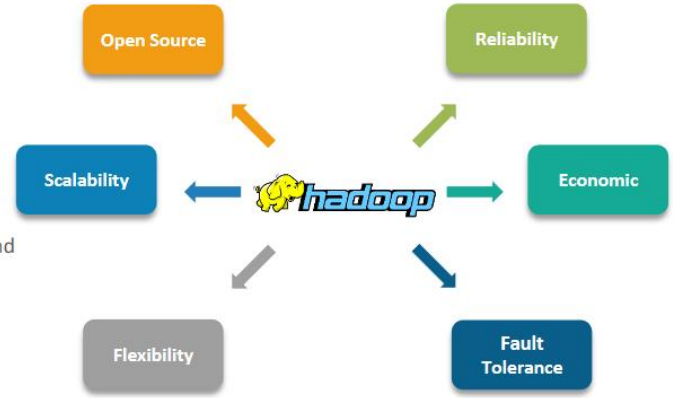
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
28 / 121
Let's start our journey with Hadoop
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 29 / 121

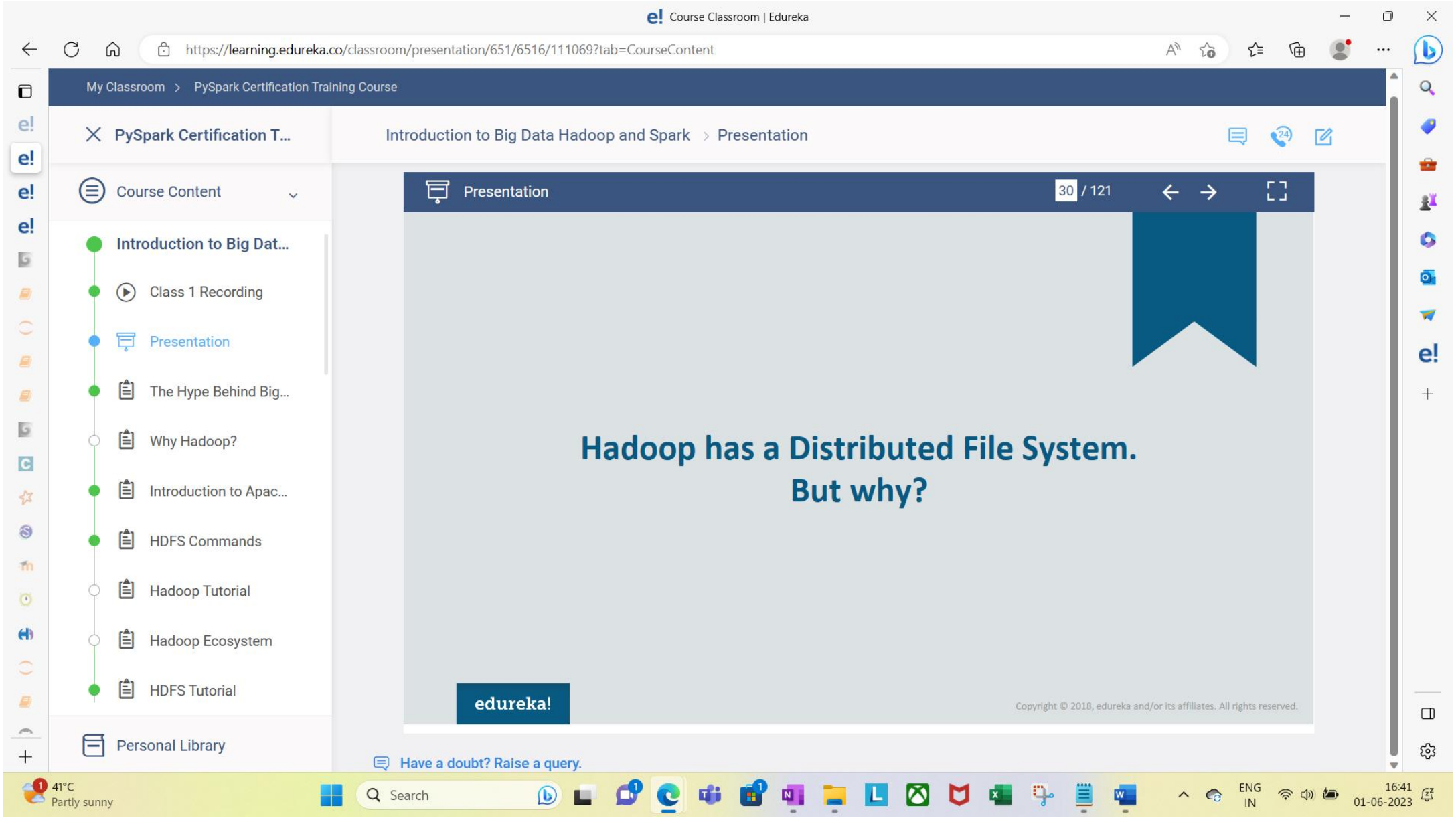
What is Hadoop?

- Apache Hadoop is an ***open-source framework*** that allows distributed processing of large data sets across clusters of commodity machines using a simple programming model
- It enables ***data management*** with ***scalable storage*** and ***distributed processing***



edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
30 / 121
Hadoop has a Distributed File System.
But why?
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 31 / 121

DFS : Distributed File System

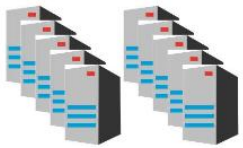
Distributed & Parallel Computation

Read 1 TB data

1 Machine

4 I/O channels
Each channel – 100 MB/s

43 Minutes



10 Machine

4 I/O channels
Each channel – 100 MB/s

Time taken = ??

edureka!

Have a doubt? Raise a query.

Presentation 32 / 121

DFS : Distributed File System

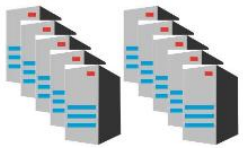
Distributed & Parallel Computation

Read 1 TB data

1 Machine

4 I/O channels
Each channel – 100 MB/s

43 Minutes



10 Machine

4 I/O channels
Each channel – 100 MB/s

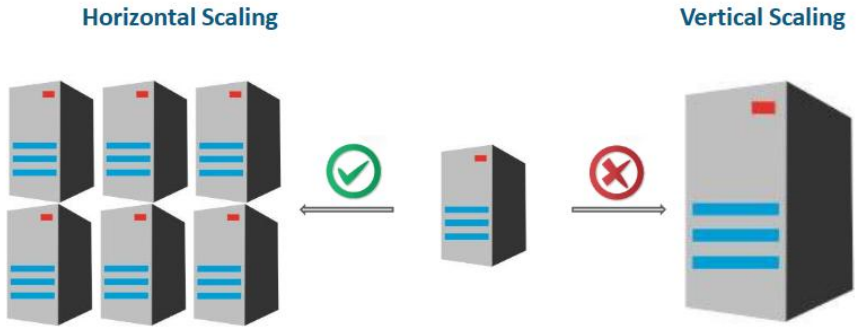
4.3 Minutes

edureka!

Have a doubt? Raise a query.

Presentation 33 / 121

DFS : Horizontal and Vertical Scaling

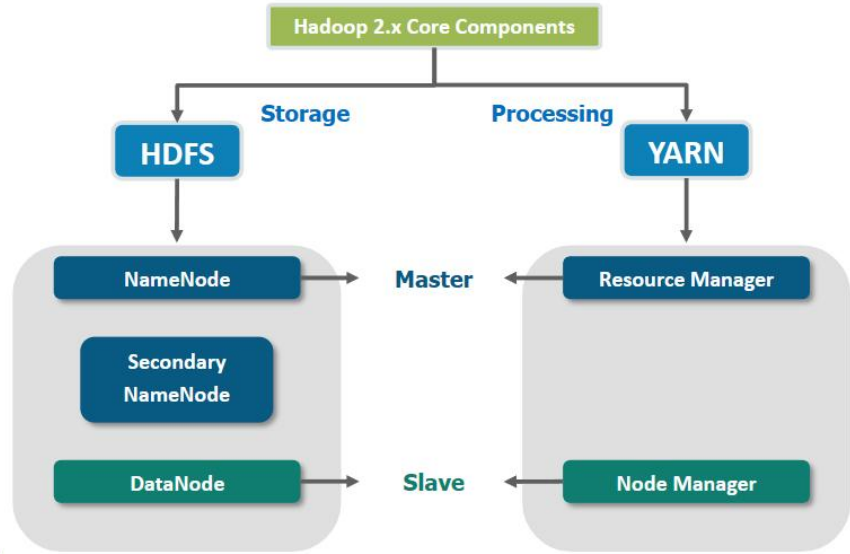


edureka!

Have a doubt? Raise a query.

Presentation 34 / 121

Hadoop 2.x Core Components



edureka!

Have a doubt? Raise a query.

Presentation 35 / 121

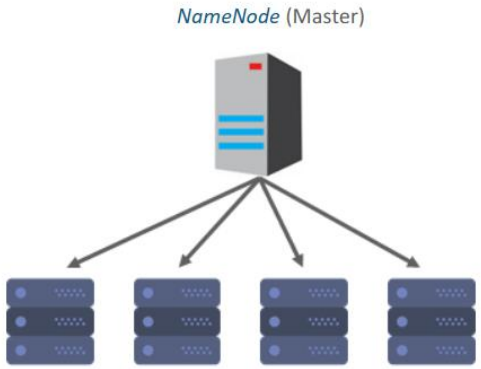
NameNode and DataNode

NameNode

- ***Master deamon*** which *maintains* and *manages* the *DataNodes* (slave nodes)
- *Records the metadata* of all the files stored in the cluster
- Regularly *receives a Heartbeat* and a *block report* from all the DataNodes in the cluster

DataNode

- ***Slave daemons*** which runs on each slave machine
- The actual *data* is *stored* on DataNodes
- Responsible for *serving read & write requests* from the clients



edureka!

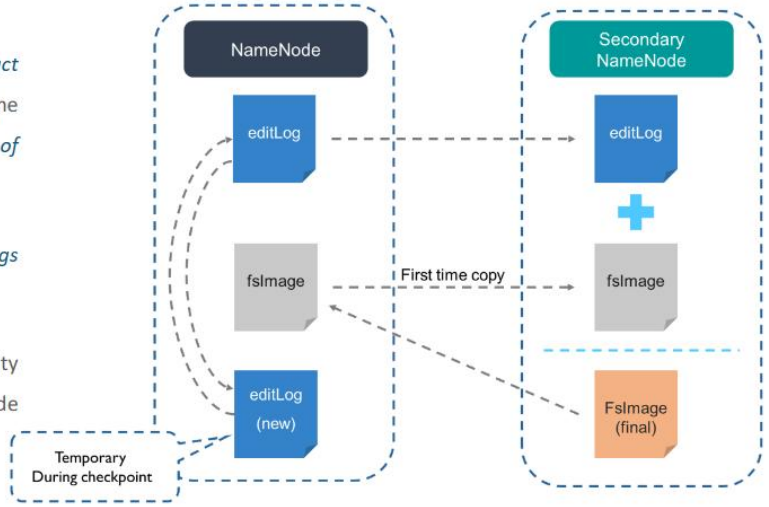
Have a doubt? Raise a query.

Presentation 36 / 121

Secondary NameNode & Checkpointing

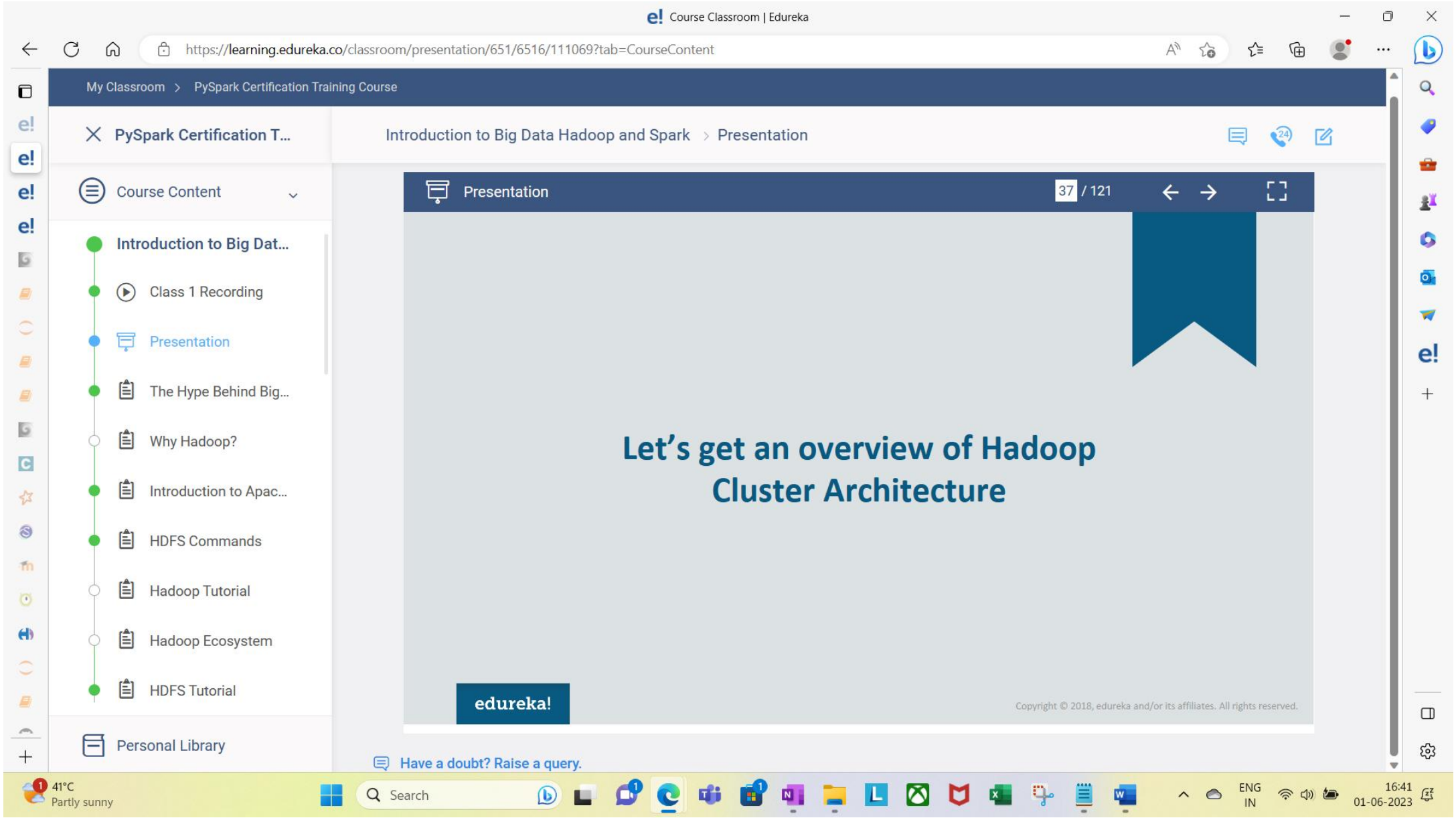
Secondary NameNode

- *Job of* ***Secondary NameNode*** is to *contact NameNode* in a *periodic manner* after certain time interval(by default 1 hour) and *pulls copy of metadata* information out of NameNode
- **Checkpointing** is a process of *combining edit logs with FsImage*
- Secondary NameNode takes over the responsibility of checkpointing, therefore making NameNode more available



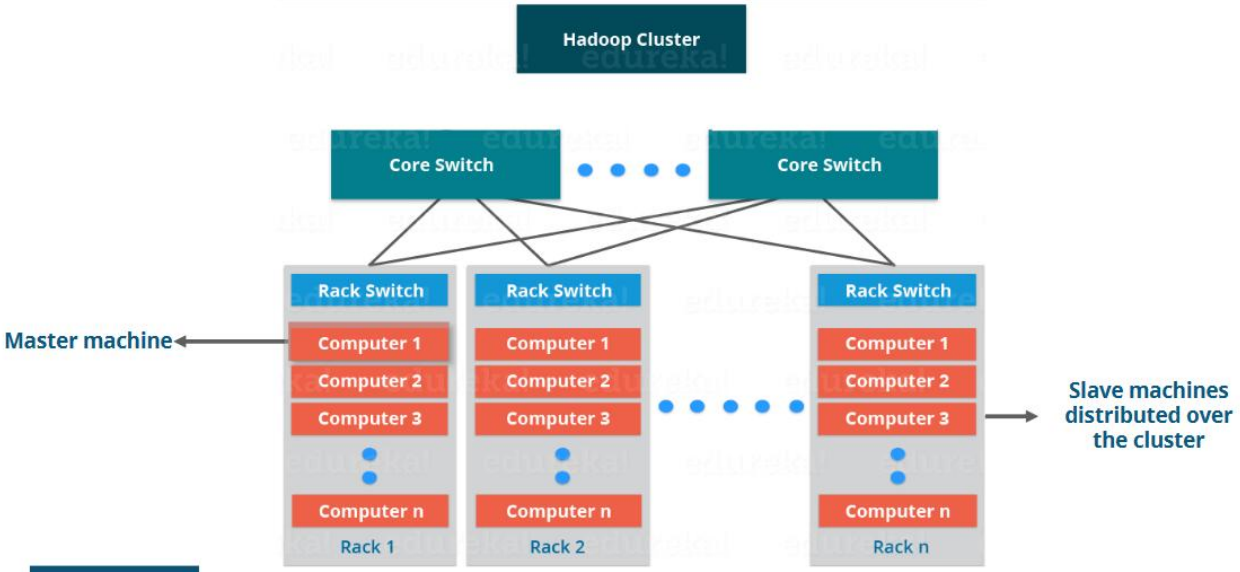
edureka!

Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
37 / 121
Let's get an overview of Hadoop Cluster Architecture
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 38 / 121

Hadoop Cluster Architecture : Master Slave Topology



edureka!

Have a doubt? Raise a query.

Presentation 39 / 121

Typical Hadoop Cluster Configuration

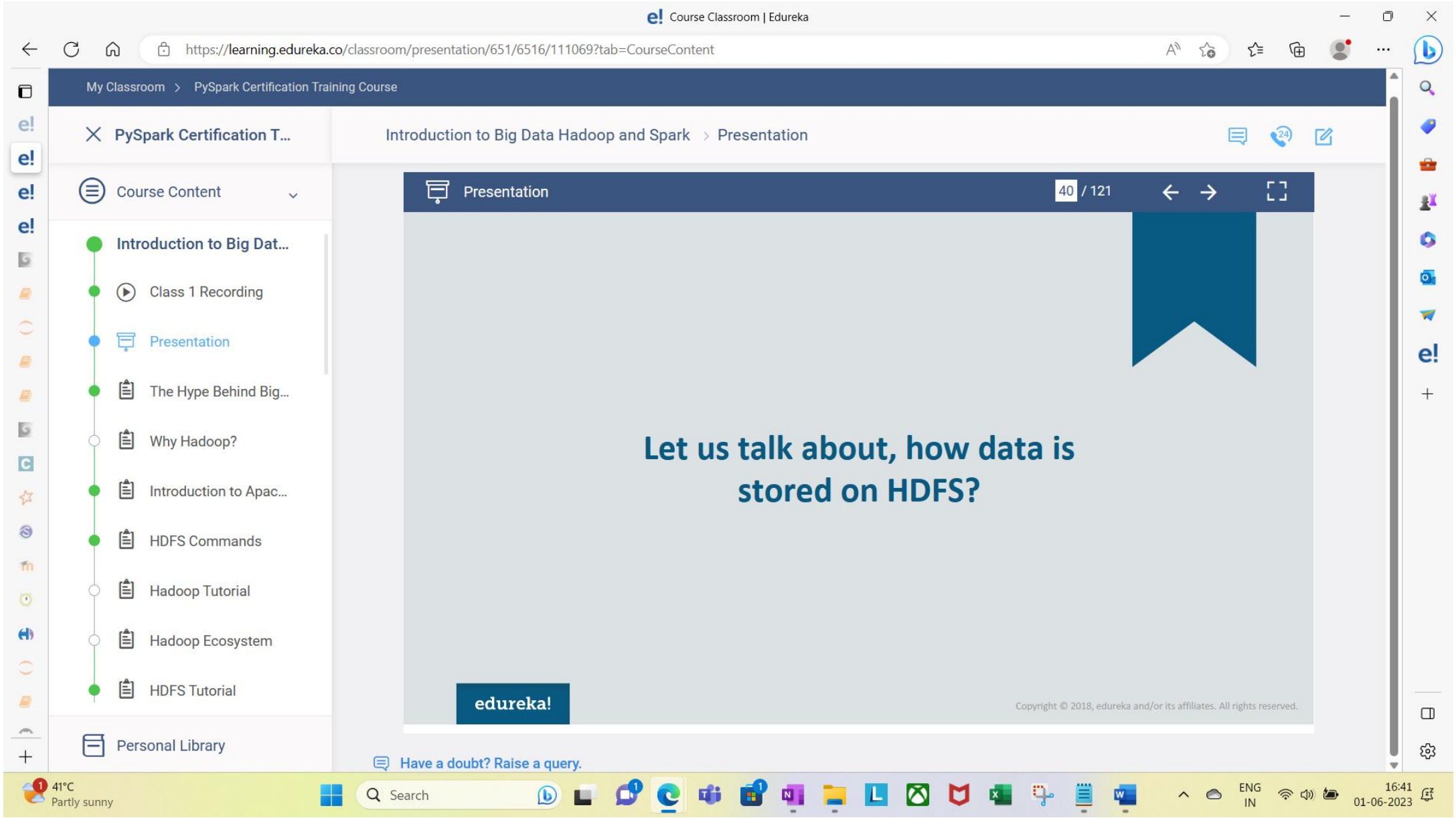
NameNode
- **RAM:** 64 GB,
- **Hard disk:** 1 TB
- **Processor:** Xenon with 8 Cores
- **Ethernet:** 3 x 10 GB/s
- **OS:** 64-bit CentOS / Linux
- **Power:** Redundant Power Supply

Secondary NameNode
- **RAM:** 32 GB
- **Hard disk:** 1 TB
- **Processor:** Xenon with 4 Cores
- **Ethernet:** 3 x 10 GB/s
- **OS:** 64-bit CentOS / Linux
- **Power:** Redundant Power Supply

DataNode
- **RAM:** 16GB
- **Hard disk:** 6 x 2TB
- **Processor:** Xenon with 2 cores
- **Ethernet:** 3 x 10 GB/s
- **OS:** 64-bit CentOS / Linux

edureka!

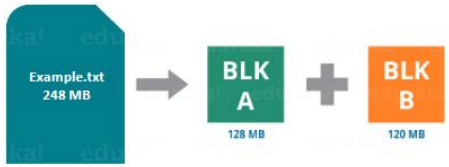
Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
40 / 121
Let us talk about, how data is stored on HDFS?
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 41 / 121

HDFS Blocks

- Each *file* is *stored* on *HDFS* as *blocks*
- The *default size* of each block is *128 MB* in Apache Hadoop 2.x
- Let us say, I have a *file "example.txt" of size 248 MB*. Below is the representation of how it will be stored on HDFS



How many blocks will be created if a file of size 514 MB is copied to HDFS ?

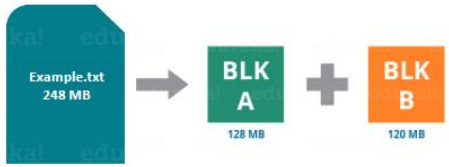
edureka!

Have a doubt? Raise a query.

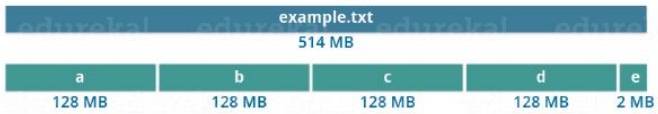
Presentation 42 / 121

HDFS Blocks

- Each *file* is *stored* on *HDFS* as *blocks*
- The *default size* of each block is *128 MB* in Apache Hadoop 2.x
- Let us say, I have a *file "example.txt" of size 248 MB*. Below is the representation of how it will be stored on HDFS

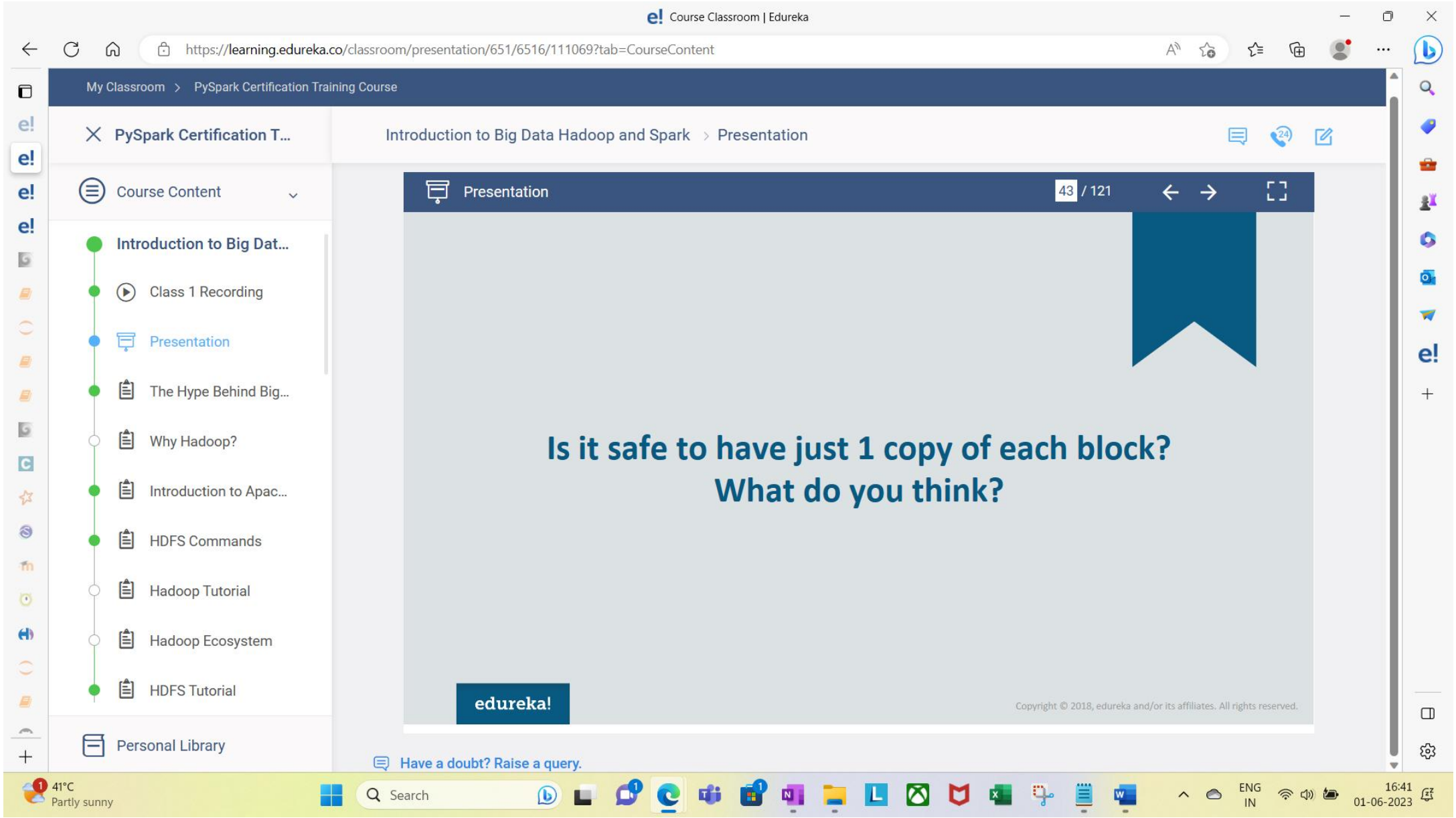


How many blocks will be created if a file of size 514 MB is copied to HDFS ?



edureka!

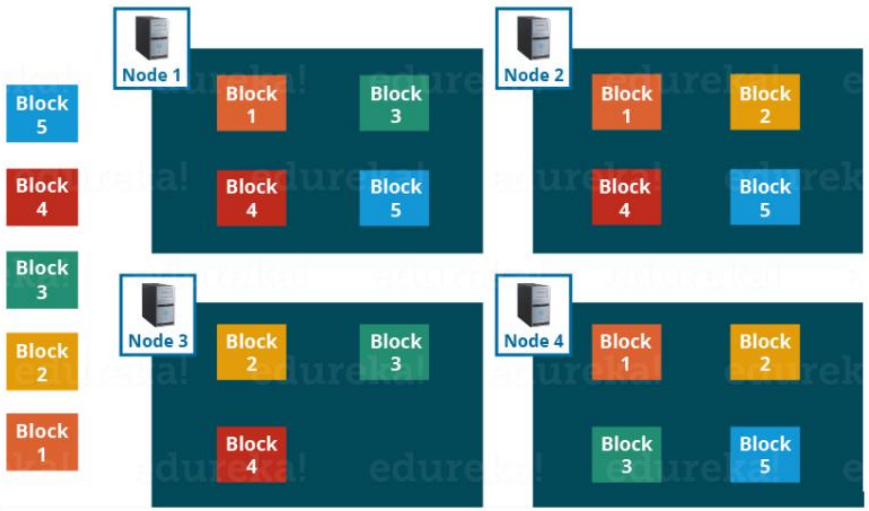
Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
43 / 121
Is it safe to have just 1 copy of each block?
What do you think?
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Partly sunny
Search
ENG
IN
16:41
01-06-2023

Presentation 44 / 121

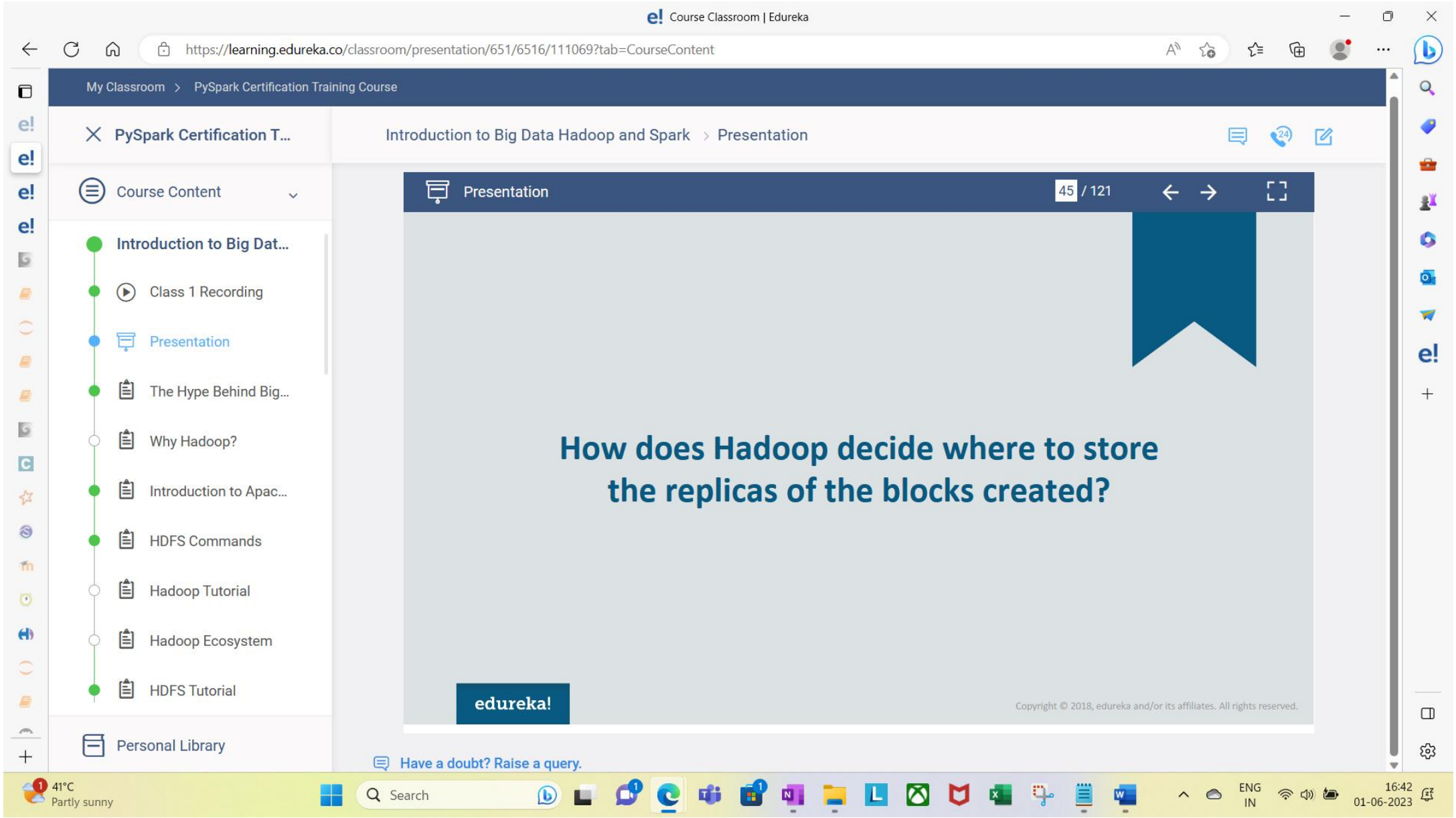
Block Replication

By Default the replication factor is **3**



edureka!

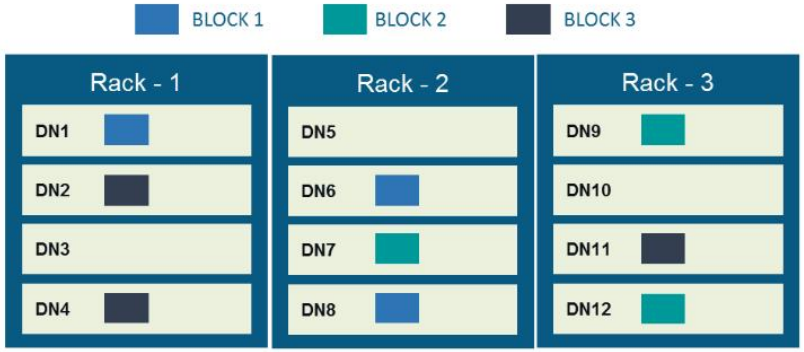
Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
45 / 121
How does Hadoop decide where to store the replicas of the blocks created?
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Partly sunny
Search
ENG
IN
16:42
01-06-2023

Presentation 46 / 121

Rack Awareness

- *Rack awareness algorithm reduces latency* as well as *provide fault tolerance* by replicating data block
- Rack awareness algorithm says that the *first replica of a block will be stored on a local rack & the next two replicas will be stored on a different (remote) rack*



edureka!

Have a doubt? Raise a query.

Presentation 47 / 121

Rack Awareness

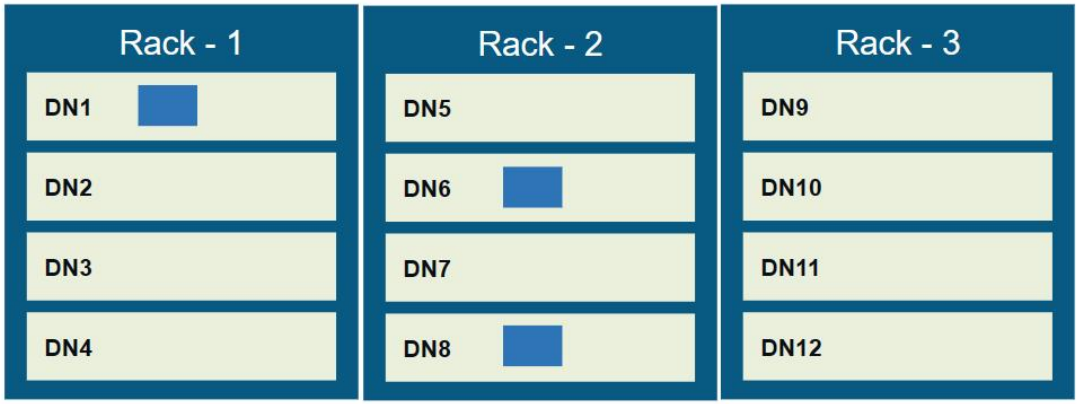
Rack - 1	Rack - 2	Rack - 3
DN1	DN5	DN9
DN2	DN6	DN10
DN3	DN7	DN11
DN4	DN8	DN12

edureka!

Have a doubt? Raise a query.

Presentation 48 / 121

Rack Awareness

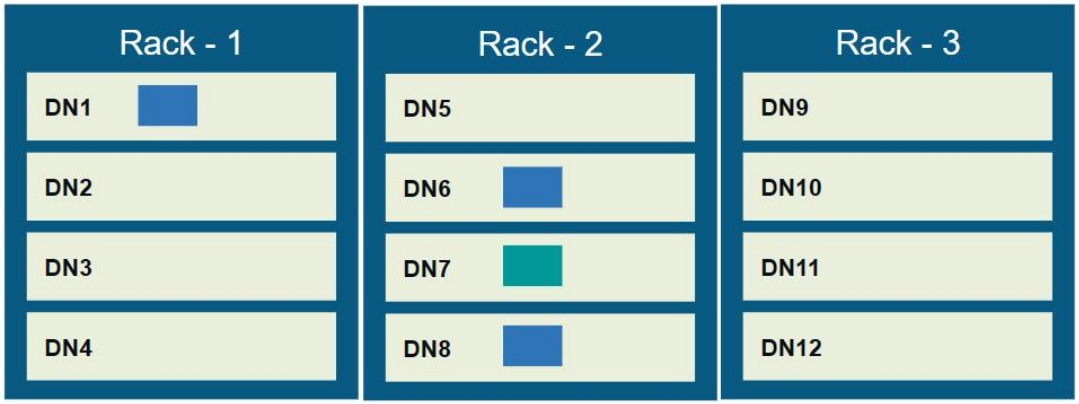


edureka!

Have a doubt? Raise a query.

Presentation 49 / 121

Rack Awareness

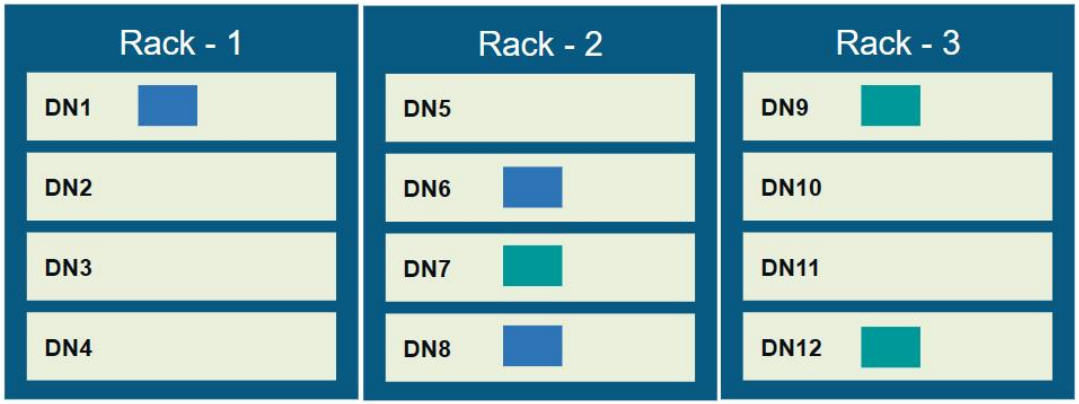


edureka!

Have a doubt? Raise a query.

Presentation 50 / 121

Rack Awareness

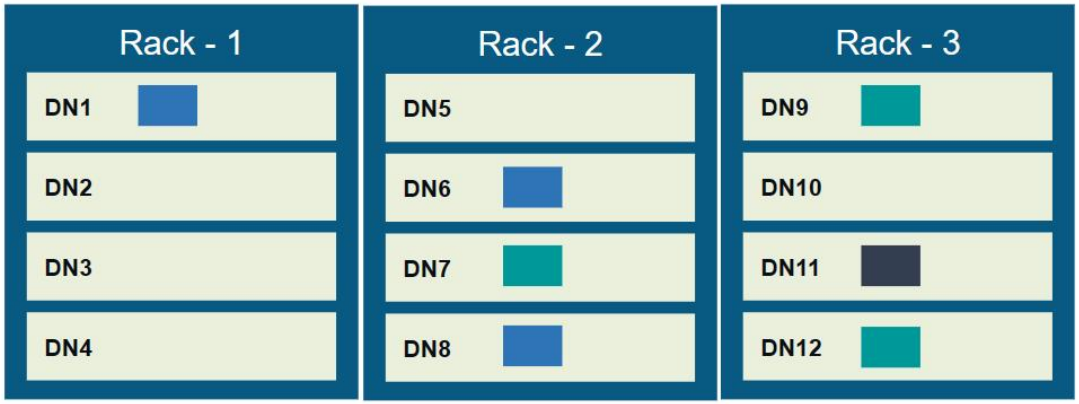


edureka!

Have a doubt? Raise a query.

Presentation 51 / 121

Rack Awareness

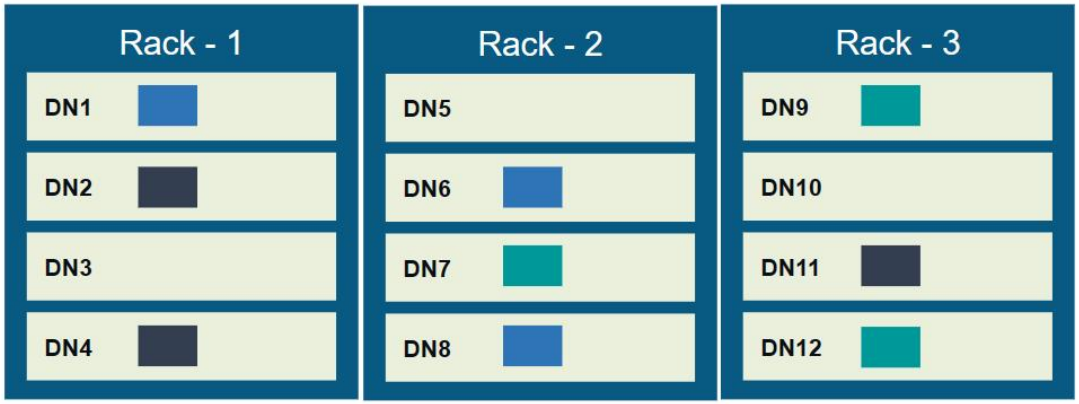


edureka!

Have a doubt? Raise a query.

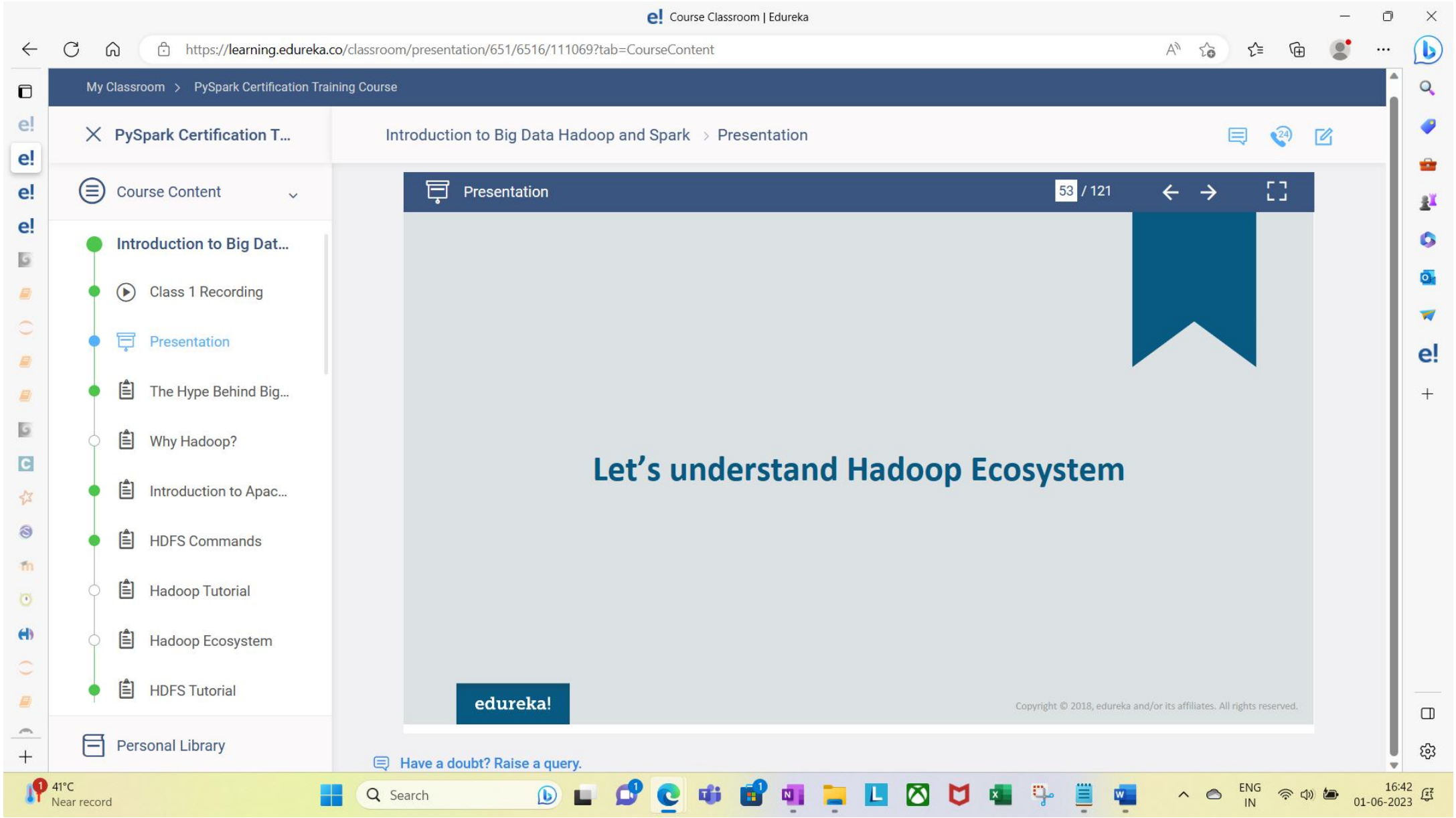
Presentation 52 / 121

Rack Awareness



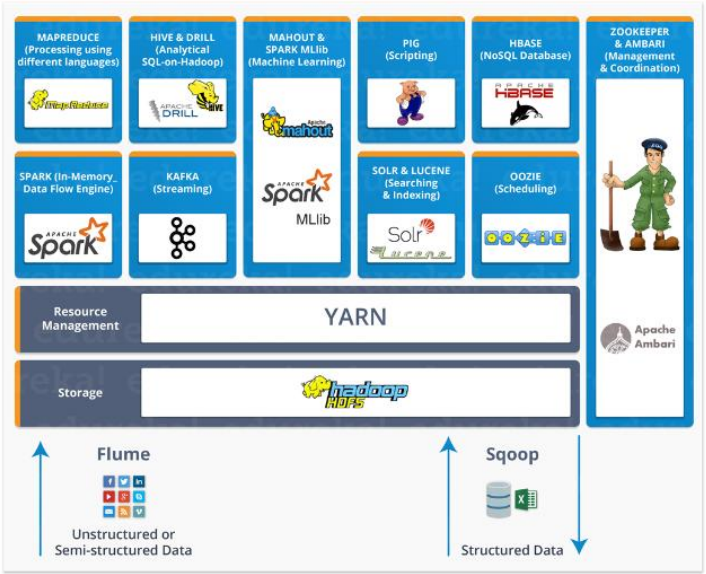
edureka!

Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
53 / 121
Let's understand Hadoop Ecosystem
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

Presentation 54 / 121

Hadoop Ecosystem



edureka!

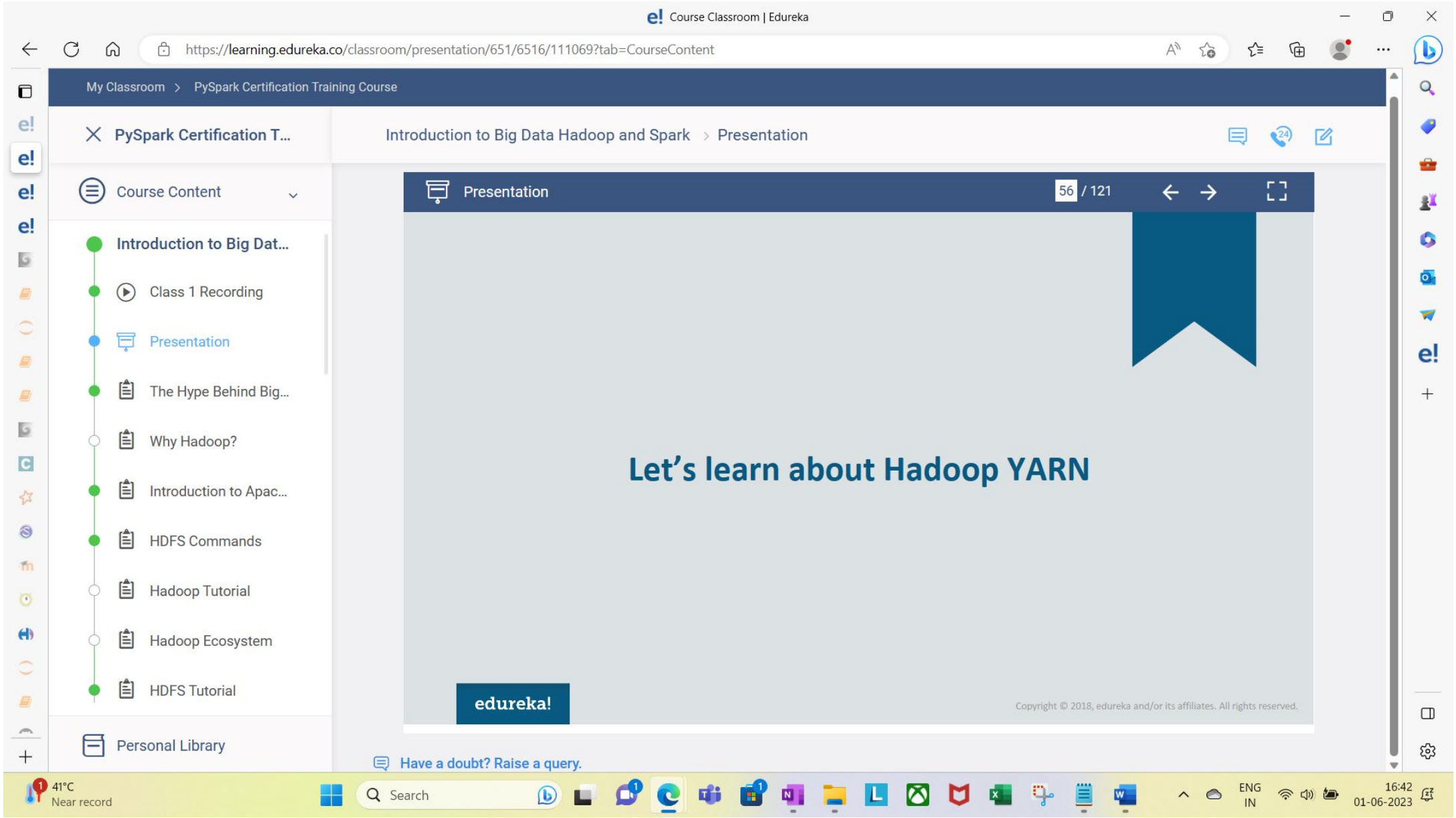
Have a doubt? Raise a query.

Components of Hadoop Ecosystem

- **Apache Hadoop** provides a ***scalable solution*** to ***store and process huge data sets*** in parallel and distributed fashion
- **Apache Spark** is an ***in-memory data processing engine*** that allows us to efficiently execute streaming, machine learning or SQL workloads
- **Apache Flume** is a ***distributed, reliable*** and ***available service*** for ***efficiently collecting, aggregating***, and ***moving large amounts of log data***
- **Apache Sqoop** is a tool designed for efficiently ***transferring bulk data*** between ***Apache Hadoop*** and ***structured datastores*** such as ***RDBMS***
- **Apache Kafka** is a ***fast, scalable, durable*** and ***fault-tolerant publish-subscribe messaging system***

edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
56 / 121
Let's learn about Hadoop YARN
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

YARN Use Case - YAHOO

YAHOO!

- Yahoo has close to 35,000 nodes running Apache Hadoop
- Performs 80,000 MapReduce jobs per day
- Over 150 PB of storage

YARN at *Yahoo* helped them increase the load on the *most heavily used Hadoop cluster*. Now performs *125,000 jobs a day* when compared to *80,000 jobs a day* which is close to *50% increase*

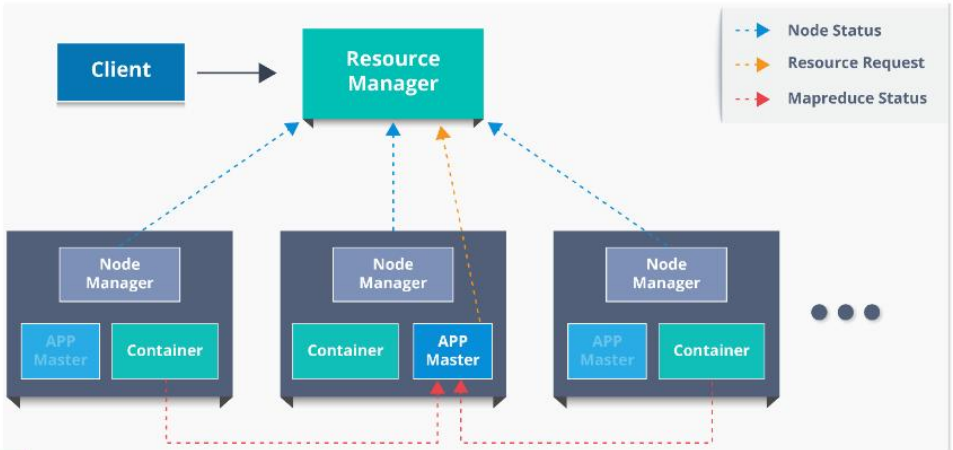
edureka!

Have a doubt? Raise a query.

Presentation 58 / 121

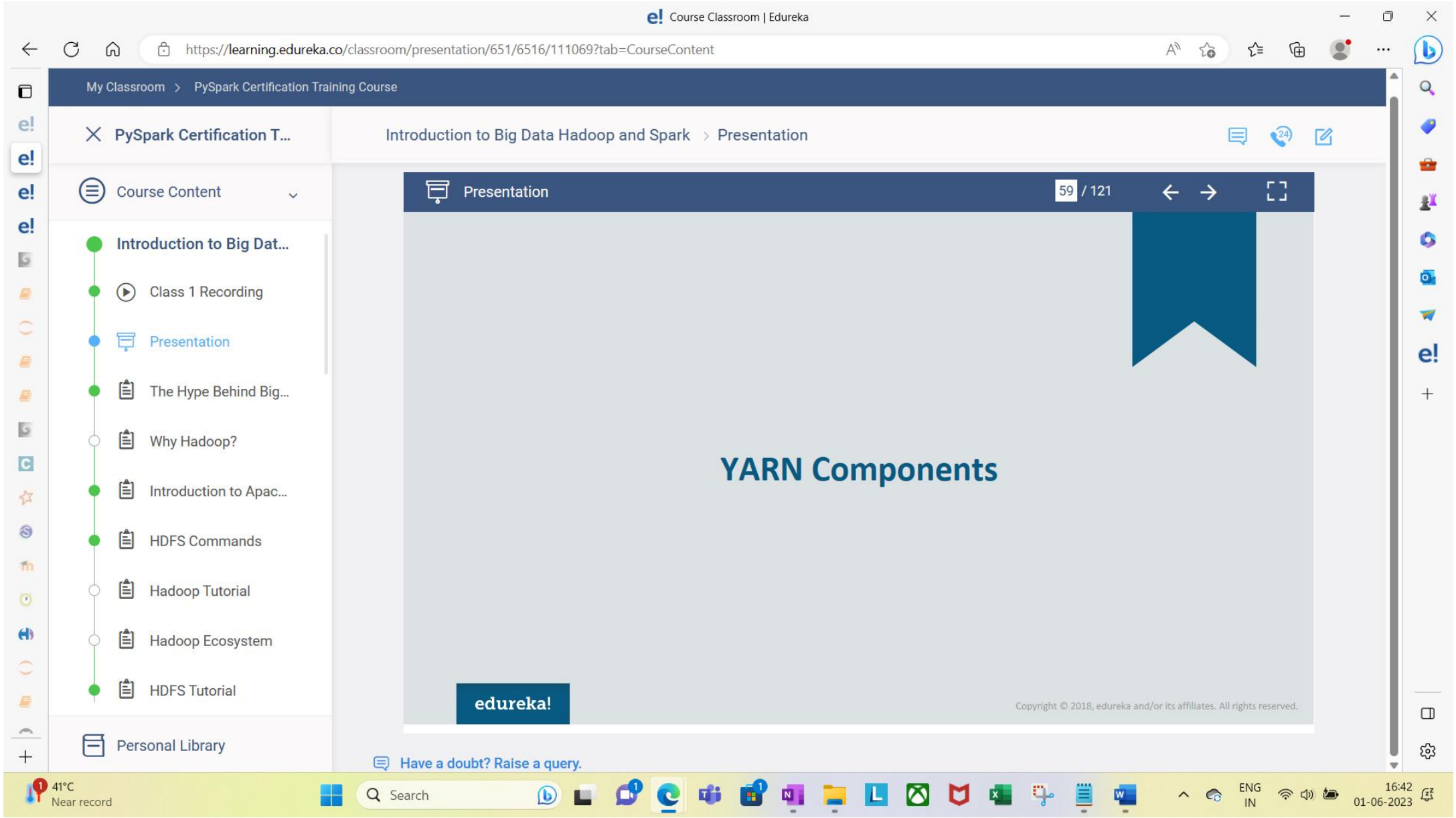
What is YARN?

- Hadoop 2.0 came up with a new framework ***YARN (Yet Another Resource Negotiator)***, which *provides ability to run Non-MapReduce application*
- YARN framework is *responsible for* doing *Cluster Resource Management*



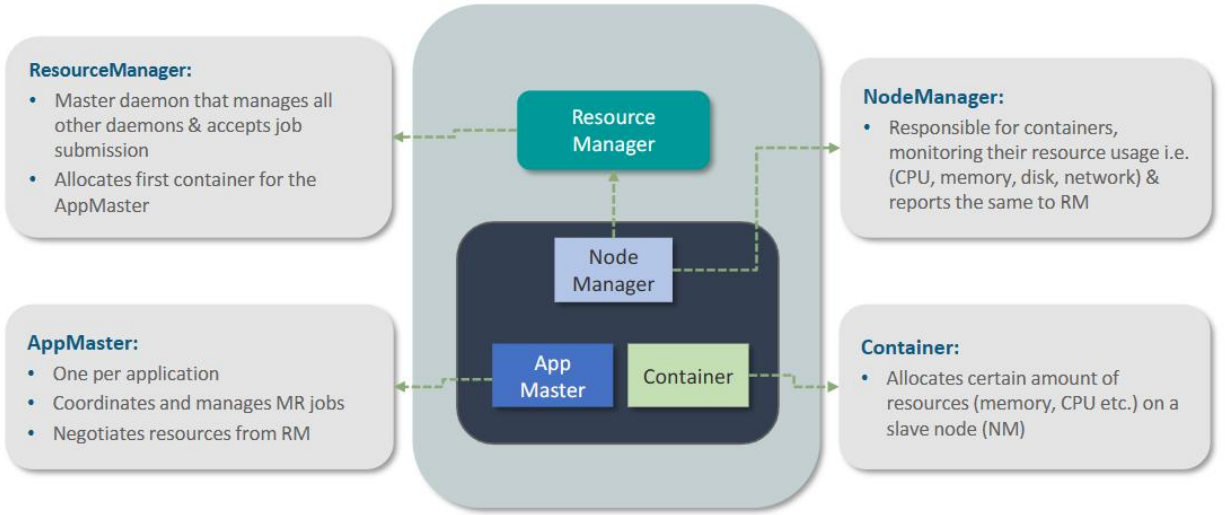
edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
59 / 121
YARN Components
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

Presentation 60 / 121

Main Components of YARN



edureka!

Have a doubt? Raise a query.

Presentation 61 / 121

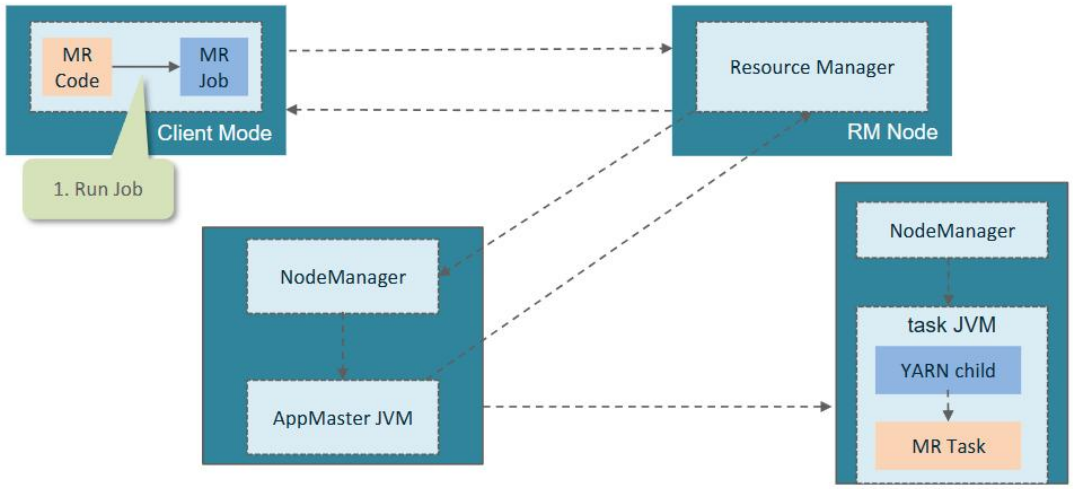
Application Submission in YARN

edureka!

Have a doubt? Raise a query.

Presentation 62 / 121

Application Submission in YARN

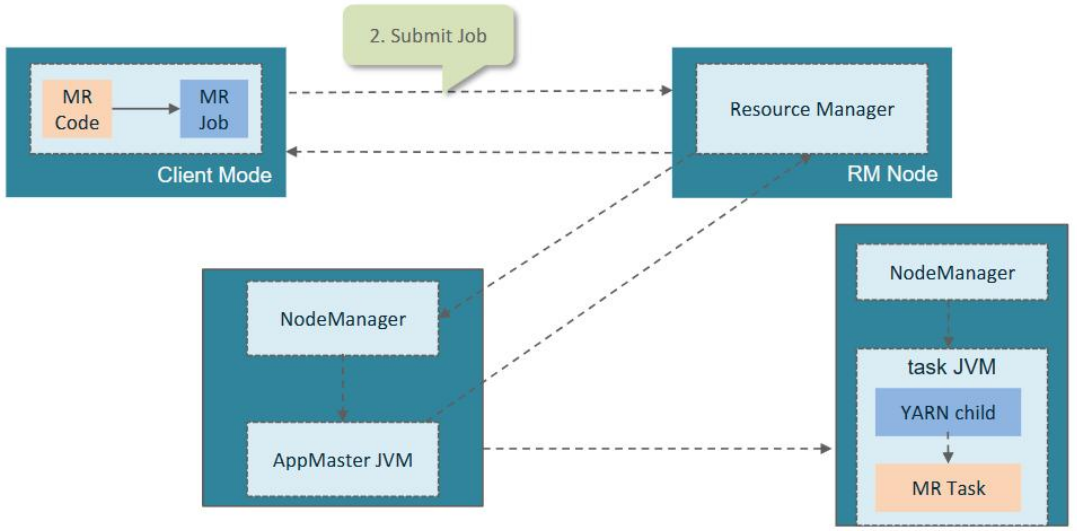


edureka!

Have a doubt? Raise a query.

Presentation 63 / 121

Application Submission in YARN

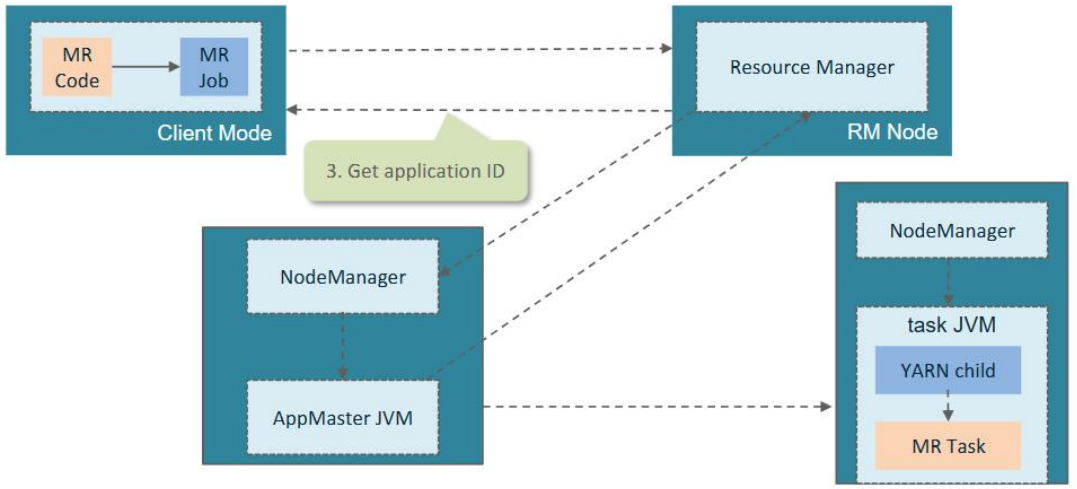


edureka!

Have a doubt? Raise a query.

Presentation 64 / 121

Application Submission in YARN

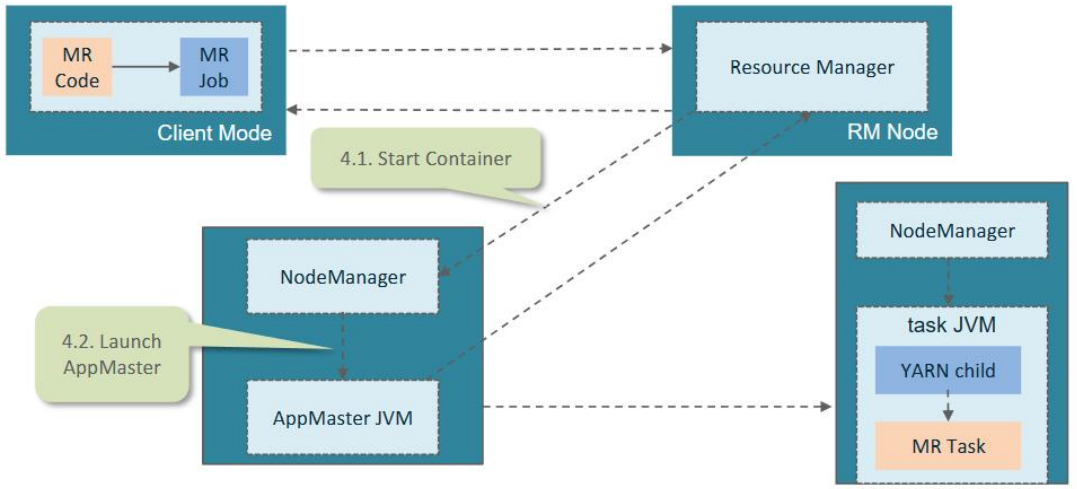


edureka!

Have a doubt? Raise a query.

Presentation 65 / 121

Application Submission in YARN

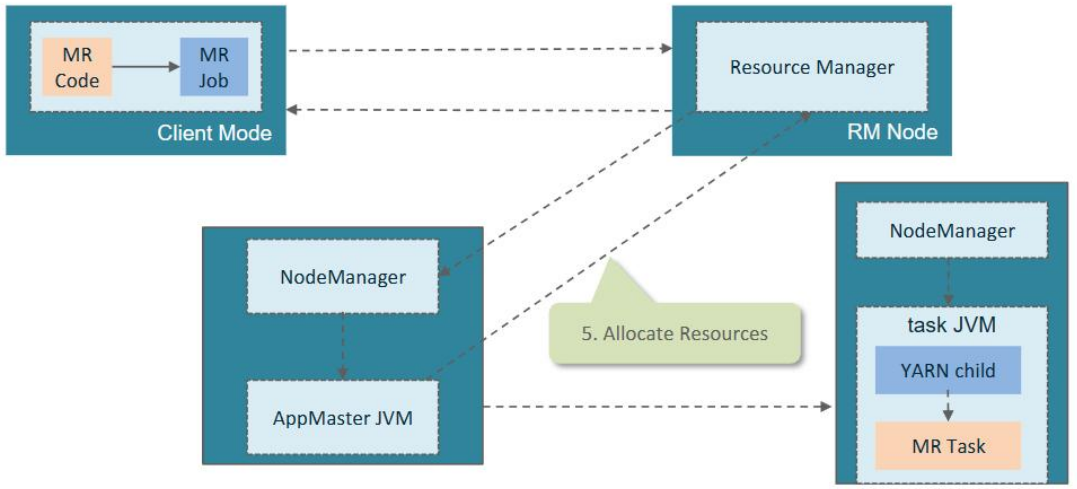


edureka!

Have a doubt? Raise a query.

Presentation 66 / 121

Application Submission in YARN

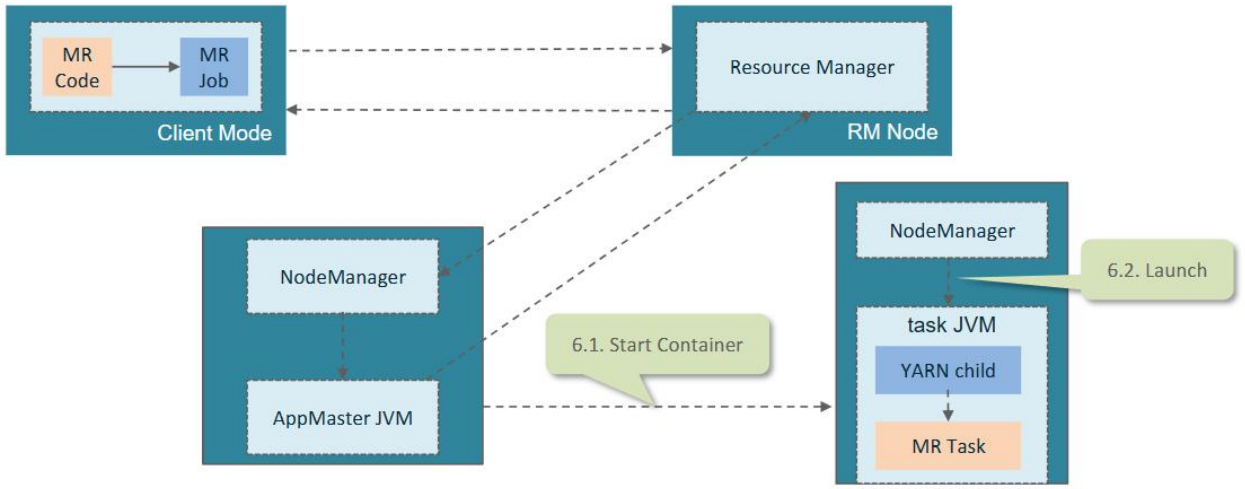


edureka!

Have a doubt? Raise a query.

Presentation 67 / 121

Application Submission in YARN

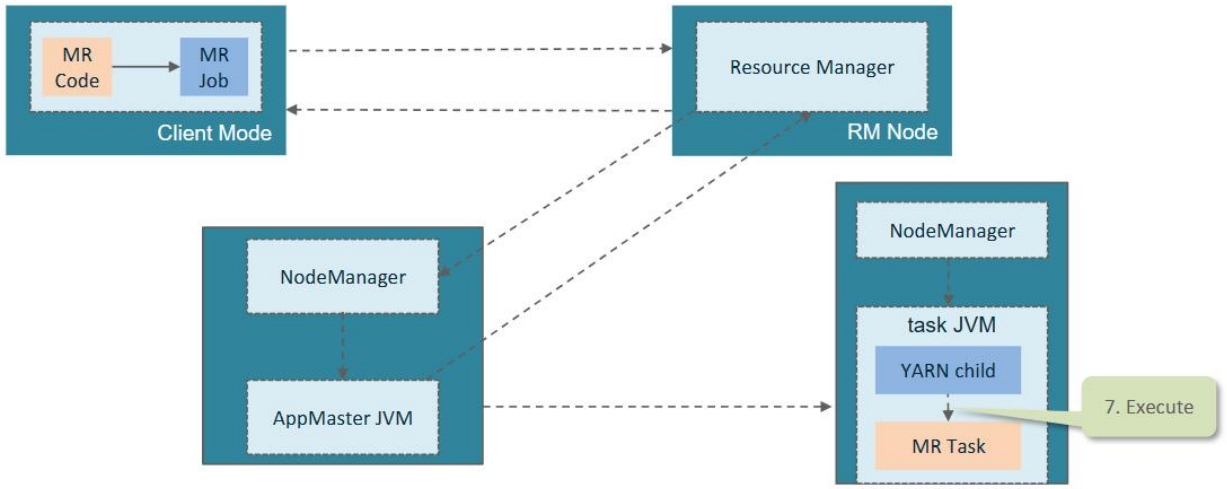


edureka!

Have a doubt? Raise a query.

Presentation 68 / 121

Application Submission in YARN



edureka!

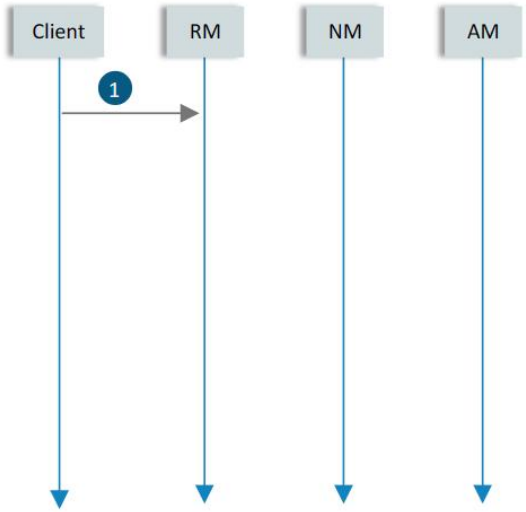
Have a doubt? Raise a query.

Presentation 69 / 121

YARN Application Workflow

➢ **Execution Sequence :**

1. Client submits an application



edureka!

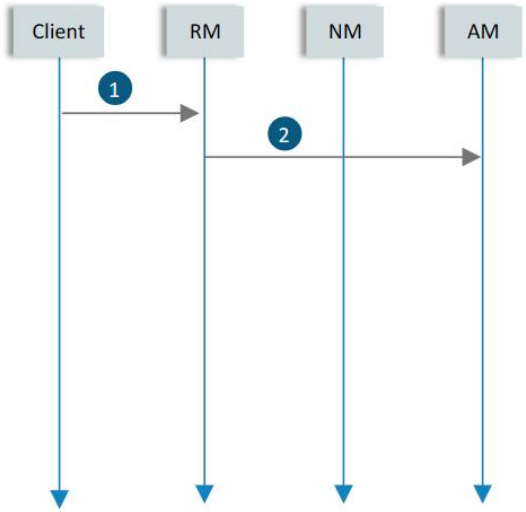
Have a doubt? Raise a query.

Presentation 70 / 121

YARN Application Workflow

➢ **Execution Sequence :**

1. Client submits an application
2. RM allocates a container to start AM



edureka!

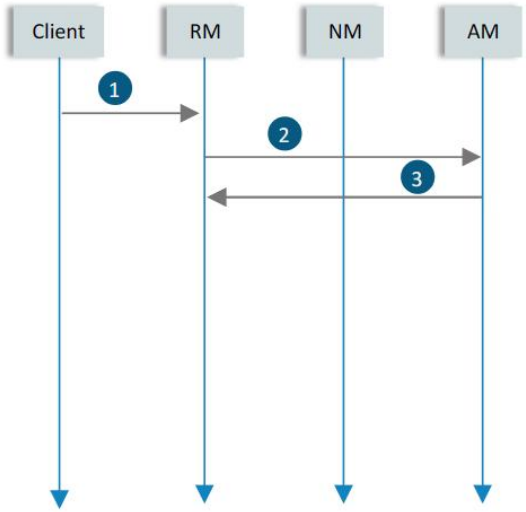
Have a doubt? Raise a query.

Presentation 71 / 121

YARN Application Workflow

➢ **Execution Sequence :**

1. Client submits an application
2. RM allocates a container to start AM
3. AM registers with RM



edureka!

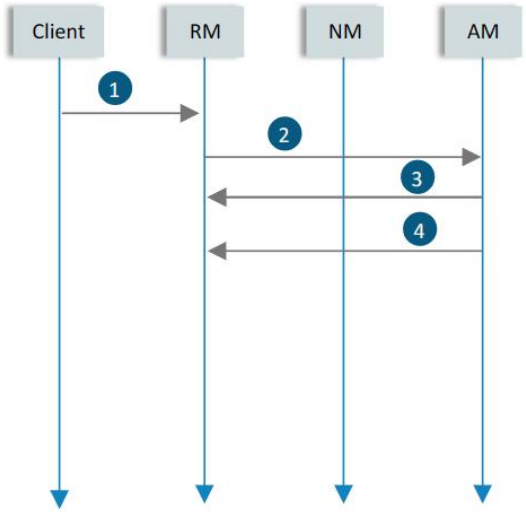
Have a doubt? Raise a query.

Presentation 72 / 121

YARN Application Workflow

Execution Sequence :

1. Client submits an application
2. RM allocates a container to start AM
3. AM registers with RM
4. AM asks containers from RM



edureka!

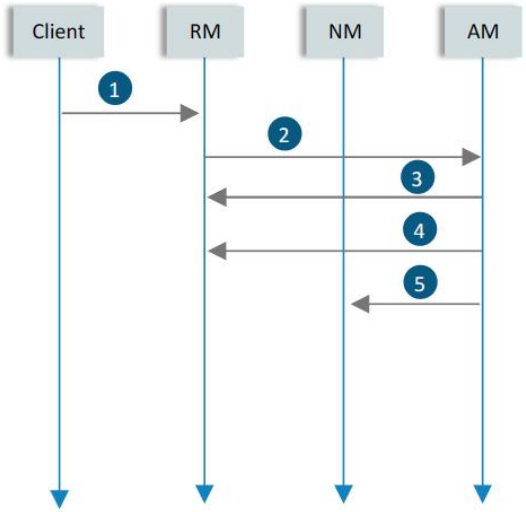
Have a doubt? Raise a query.

Presentation 73 / 121

YARN Application Workflow

Execution Sequence :

1. Client submits an application
2. RM allocates a container to start AM
3. AM registers with RM
4. AM asks containers from RM
5. AM notifies NM to launch containers



edureka!

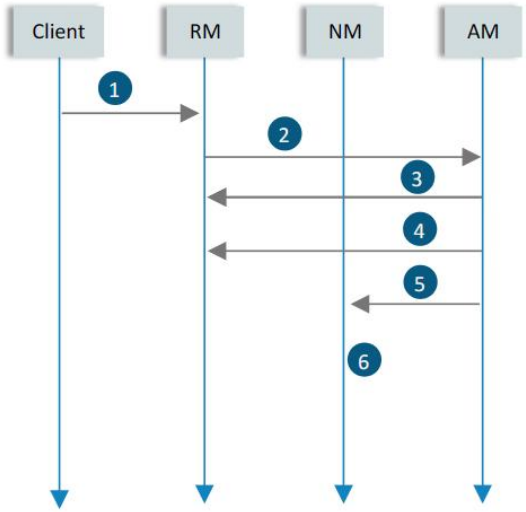
Have a doubt? Raise a query.

Presentation 74 / 121

YARN Application Workflow

Execution Sequence :

1. Client submits an application
2. RM allocates a container to start AM
3. AM registers with RM
4. AM asks containers from RM
5. AM notifies NM to launch containers
6. Application code is executed in container



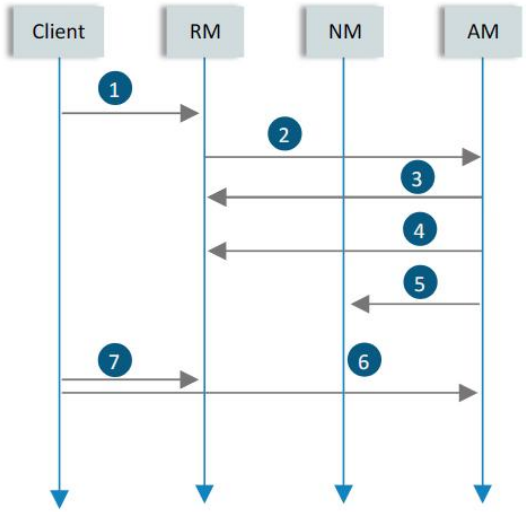
edureka!

Have a doubt? Raise a query.

YARN Application Workflow

Execution Sequence :

1. Client submits an application
2. RM allocates a container to start AM
3. AM registers with RM
4. AM asks containers from RM
5. AM notifies NM to launch containers
6. Application code is executed in container
7. Client contacts RM/AM to monitor application's status



edureka!

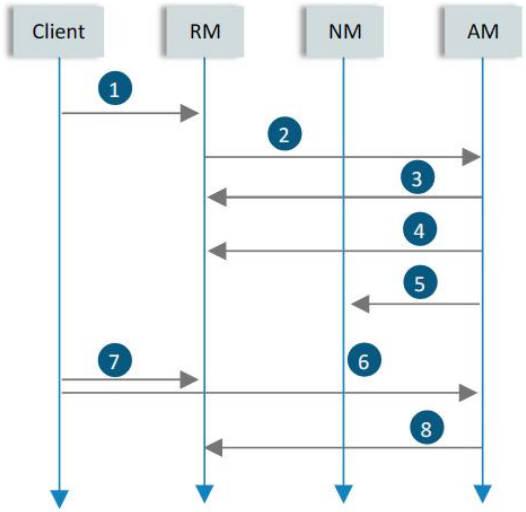
Have a doubt? Raise a query.

Presentation 76 / 121

YARN Application Workflow

Execution Sequence :

1. Client submits an application
2. RM allocates a container to start AM
3. AM registers with RM
4. AM asks containers from RM
5. AM notifies NM to launch containers
6. Application code is executed in container
7. Client contacts RM/AM to monitor application's status
8. AM unregisters with RM



edureka!

Have a doubt? Raise a query.

Presentation 77 / 121

Hadoop Cluster Modes

edureka!

Have a doubt? Raise a query.

Hadoop Cluster Modes

Standalone (or Local) Mode

- No daemons, everything runs in a single JVM
- Suitable for running MapReduce programs during development
- Has no DFS or Distributed File System

Pseudo Distributed Mode

- All Hadoop daemons run on the local machine

Multi-Node Cluster Mode

- Hadoop daemons run on a cluster of machines

edureka!

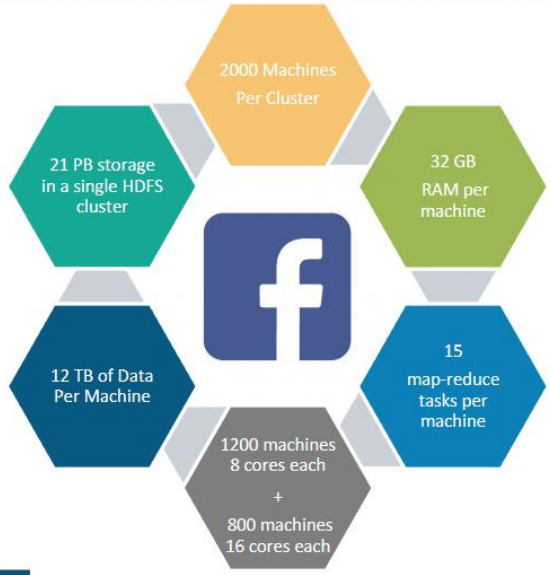
Have a doubt? Raise a query.

Presentation 79 / 121

Have a doubt? Raise a query.

Presentation 80 / 121

Hadoop Cluster : Facebook Use Case



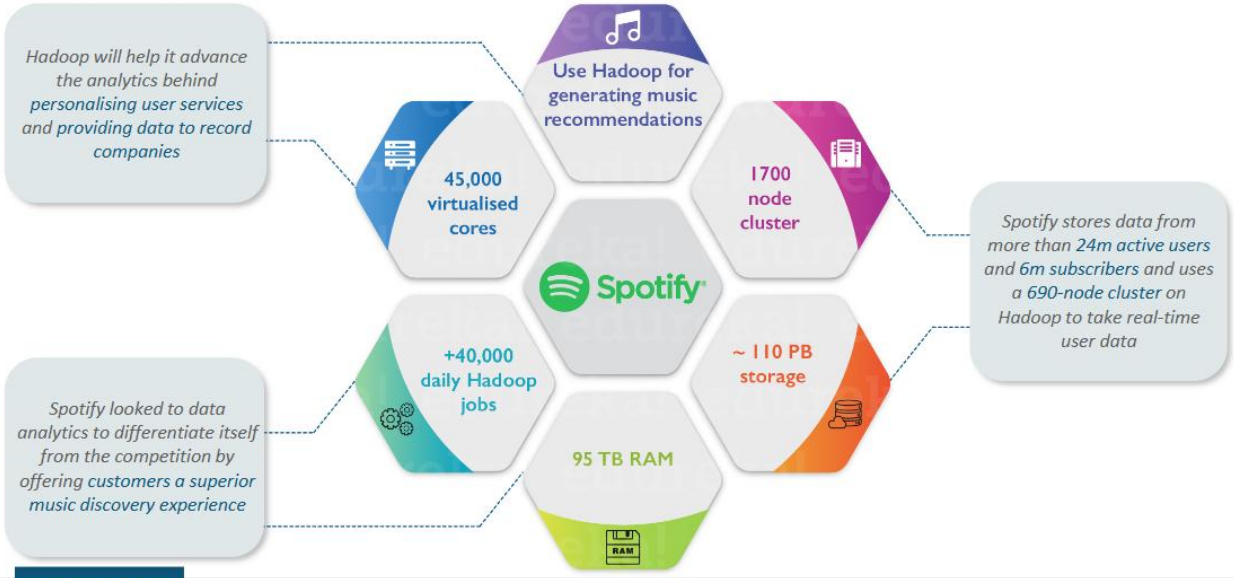
That's a total of *more than 21 PB* of *configured storage capacity*, this is larger than the previously known Yahoo!'s cluster of 14 PB

edureka!

Have a doubt? Raise a query.

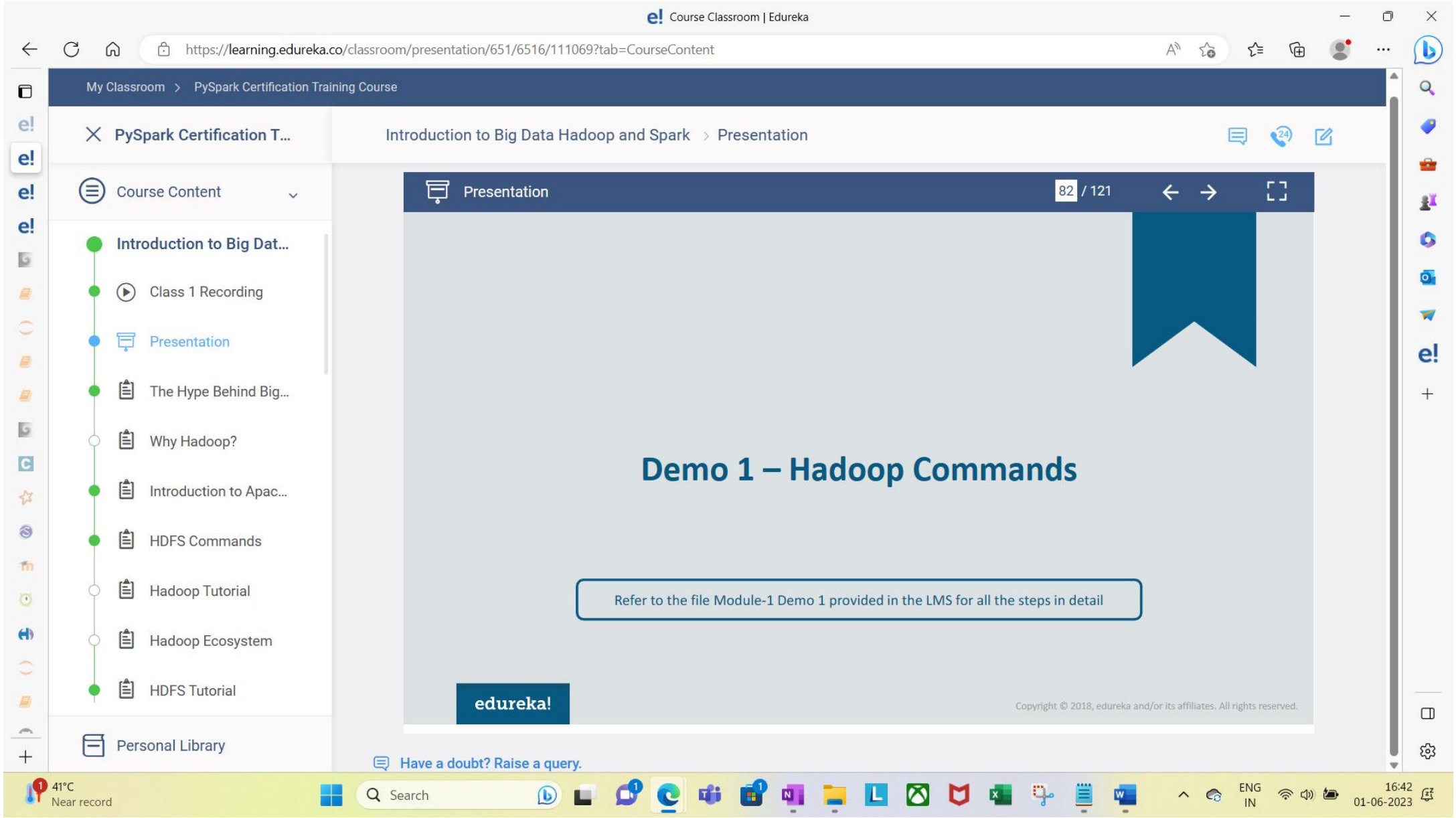
Presentation 81 / 121

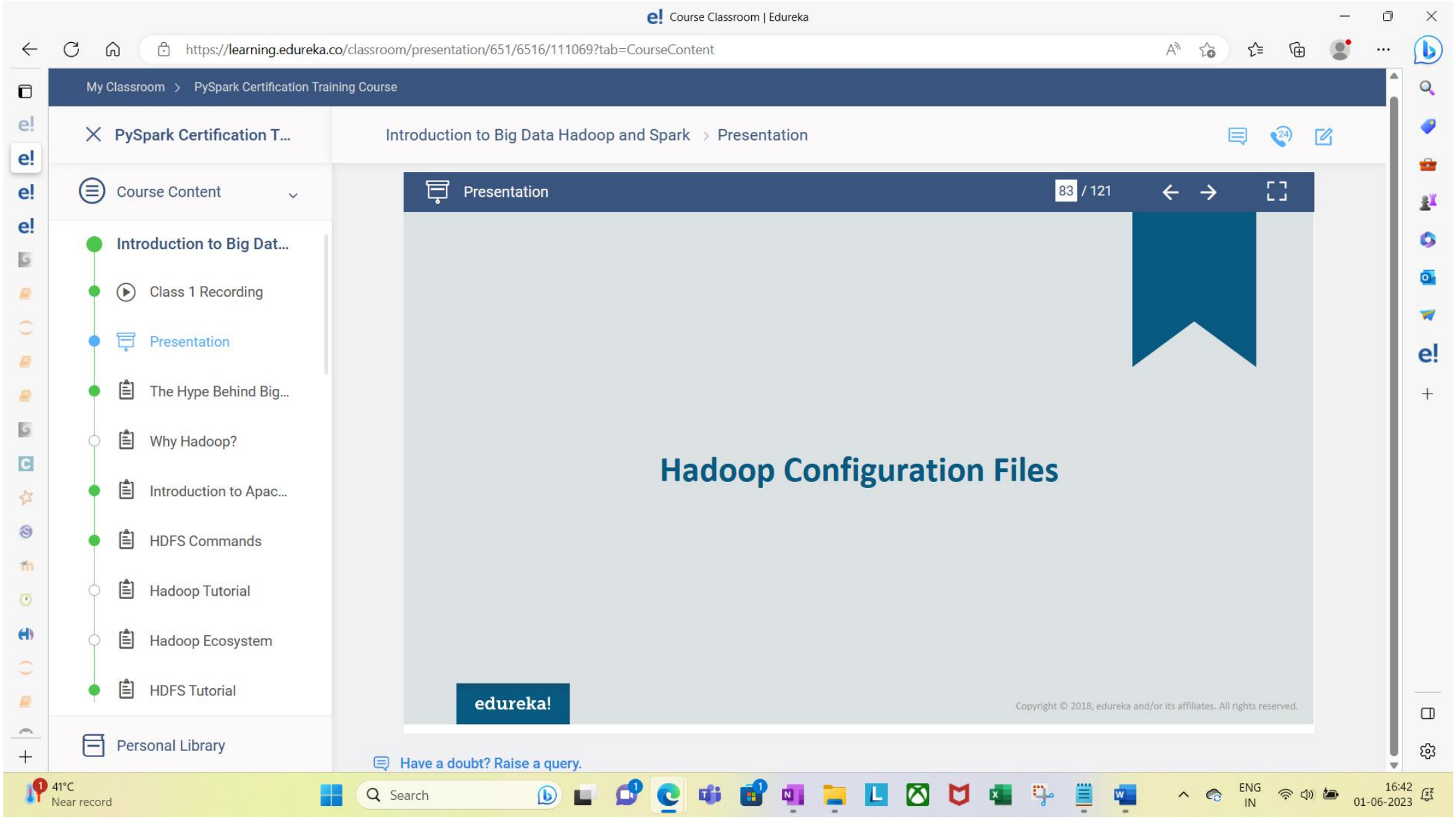
Hadoop Cluster : Spotify Use Case



edureka!

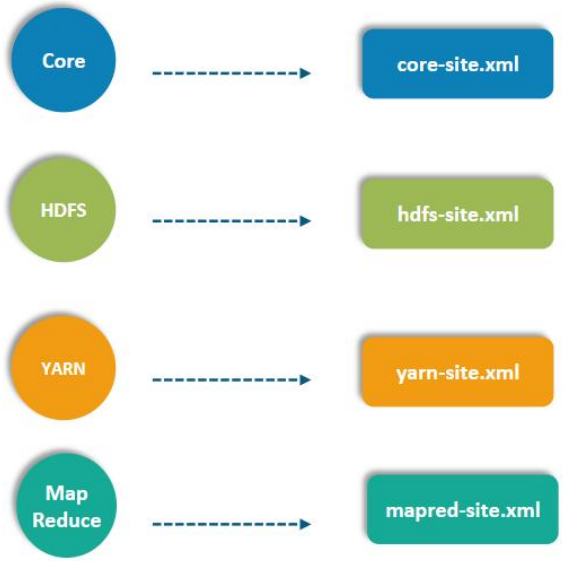
Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
82 / 121
Demo 1 – Hadoop Commands
Refer to the file Module-1 Demo 1 provided in the LMS for all the steps in detail
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
83 / 121
Hadoop Configuration Files
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

Presentation 84 / 121

Hadoop 2.X Configuration Files



edureka!

Have a doubt? Raise a query.

Hadoop 2.X Configuration Files : Description

Configuration Filenames	Description
hadoop-env.sh	Environment variables that are used in the scripts to run Hadoop.
core-site.xml	Configuration settings for Hadoop Core such as I/O settings that are common to HDFS and MapReduce.
hdfs-site.xml	Configuration settings for HDFS daemons, the namenode, the secondary namenode and the data nodes.
mapred-site.xml	Configuration settings for MapReduce Applications.
yarn-site.xml	Configuration settings for ResourceManager and NodeManager.
masters	A list of machines (one per line) that each run a secondary namenode.
slaves	A list of machines (one per line) that each run a Datanode and a NodeManager.

edureka!

Have a doubt? Raise a query.

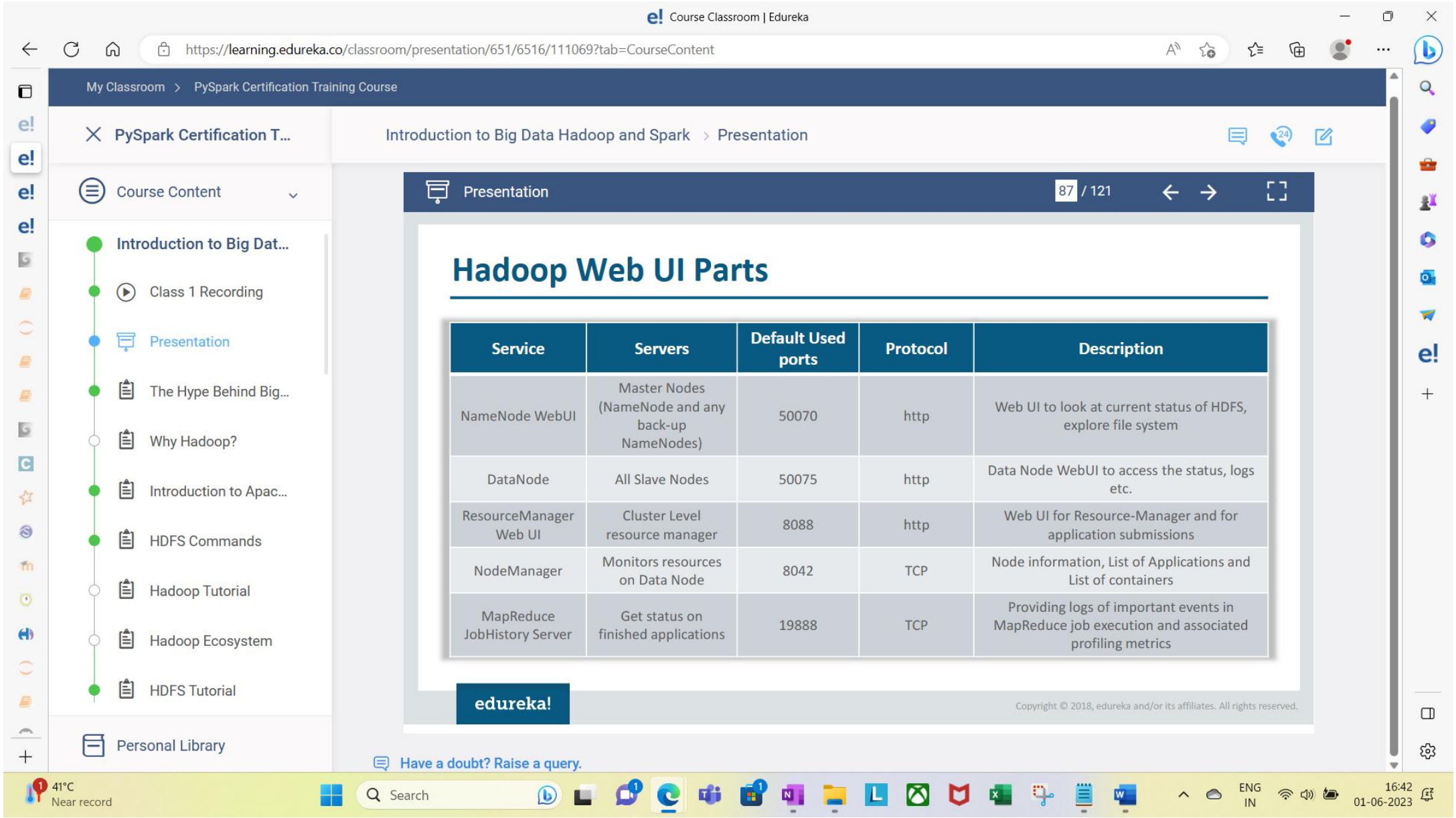
Presentation 86 / 121

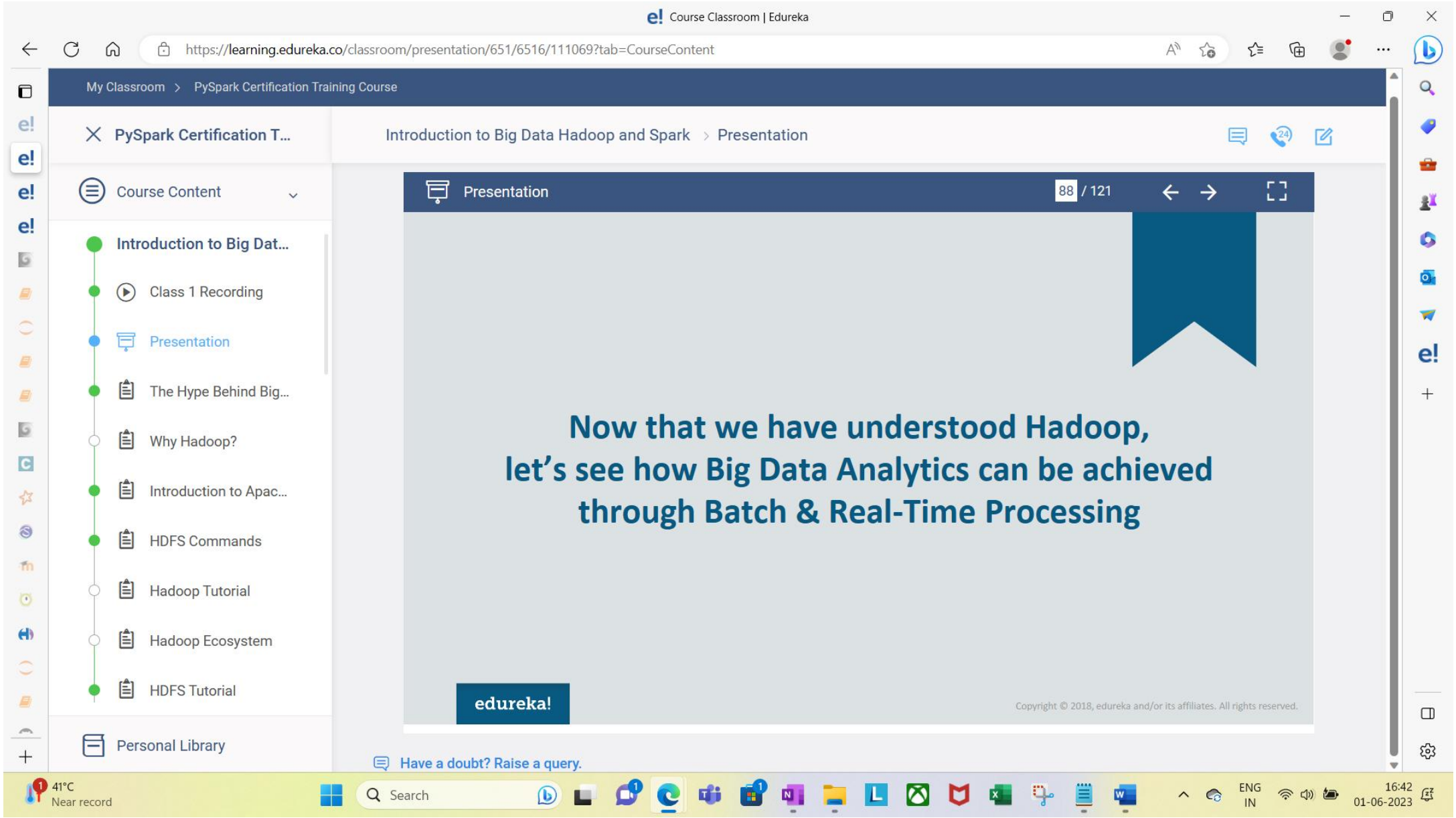
Hadoop 2.X Configuration Files

```
[edureka_249489@ip-20-0-41-93 ~]$ cd /opt/cloudera/parcels/CDH/lib/hadoop/etc/hadoop
[edureka_249489@ip-20-0-41-93 hadoop]$ ls
__cloudera_generation__  core-site.xml  hdfs-site.xml    mapred-site.xml  topology.map  yarn-site.xml
__cloudera_metadata__    hadoop-env.sh  log4j.properties  ssl-client.xml   topology.py
[edureka_249489@ip-20-0-41-93 hadoop]$
```

edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
87 / 121
Hadoop Web UI Parts
Service | Servers | Default Used ports | Protocol | Description
NameNode WebUI | Master Nodes (NameNode and any back-up NameNodes) | 50070 | http | Web UI to look at current status of HDFS, explore file system
DataNode | All Slave Nodes | 50075 | http | Data Node WebUI to access the status, logs etc.
ResourceManager Web UI | Cluster Level resource manager | 8088 | http | Web UI for Resource-Manager and for application submissions
NodeManager | Monitors resources on Data Node | 8042 | TCP | Node information, List of Applications and List of containers
MapReduce JobHistory Server | Get status on finished applications | 19888 | TCP | Providing logs of important events in MapReduce job execution and associated profiling metrics
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

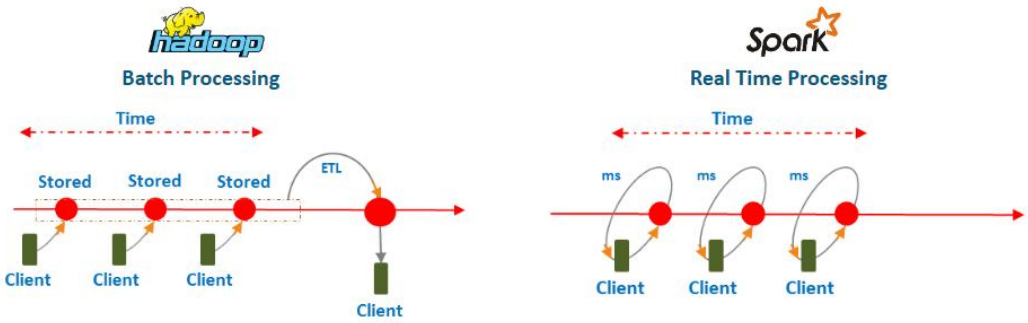

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
88 / 121
Now that we have understood Hadoop,
let's see how Big Data Analytics can be achieved
through Batch & Real-Time Processing
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

Presentation 89 / 121

Big Data Analytics

- ***Big Data Analytics*** is the *process of examining large data sets* to uncover *hidden patterns, unknown correlations, market trends, customer preferences* and other useful business information

- Analysis on Big Data can be achieved in two ways:
  - Batch processing
  - Real-Time processing



edureka!

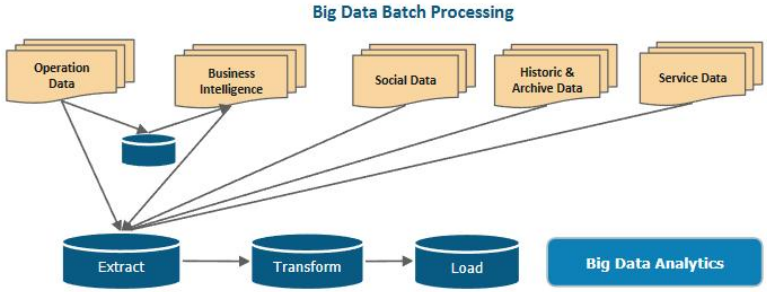
Have a doubt? Raise a query.

Presentation 90 / 121

Batch Processing

- Preferred way for processing high volumes of data
- Runs over ***group of transactions***, collected after a ***certain time period***
- Data is ***collected***, ***entered***, ***processed for every batch*** window

  (***Hadoop*** is focused on batch data processing)
- Batch processing requires separate programs for input, process and output
- Examples:
  - Canned reports
  - Interest accruals processing
  - Forecasting
  - Customers profiling



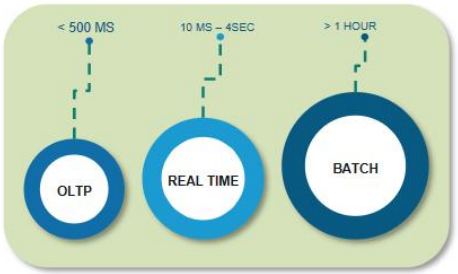
edureka!

Have a doubt? Raise a query.

Presentation 91 / 121

Real-Time Processing

- It involves a ***continual input***, process and output of data
- Data processing time is much smaller (***in fractions of seconds***)
- One such example is a ***complex event processing*** (CEP combines data from multiple sources to infer events or patterns that suggest more complicated circumstances) platform
- Another example is ***operational intelligence*** (OI is a form of real-time dynamic, business analytics that delivers visibility and insight into business operations) platforms
- There are continuously running programs which keep consuming data from streams
- They keep running forever, till they are manually stopped

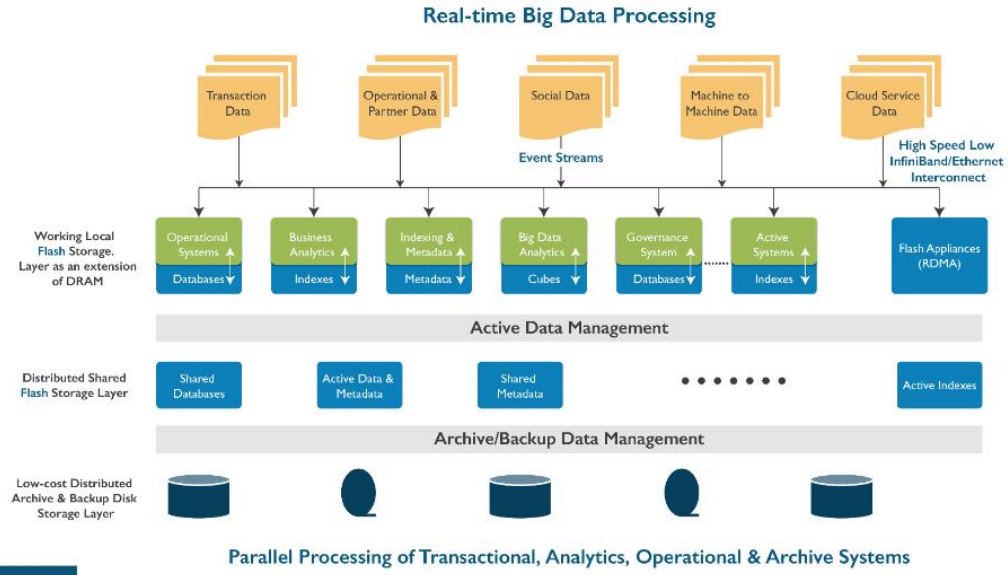


edureka!

Have a doubt? Raise a query.

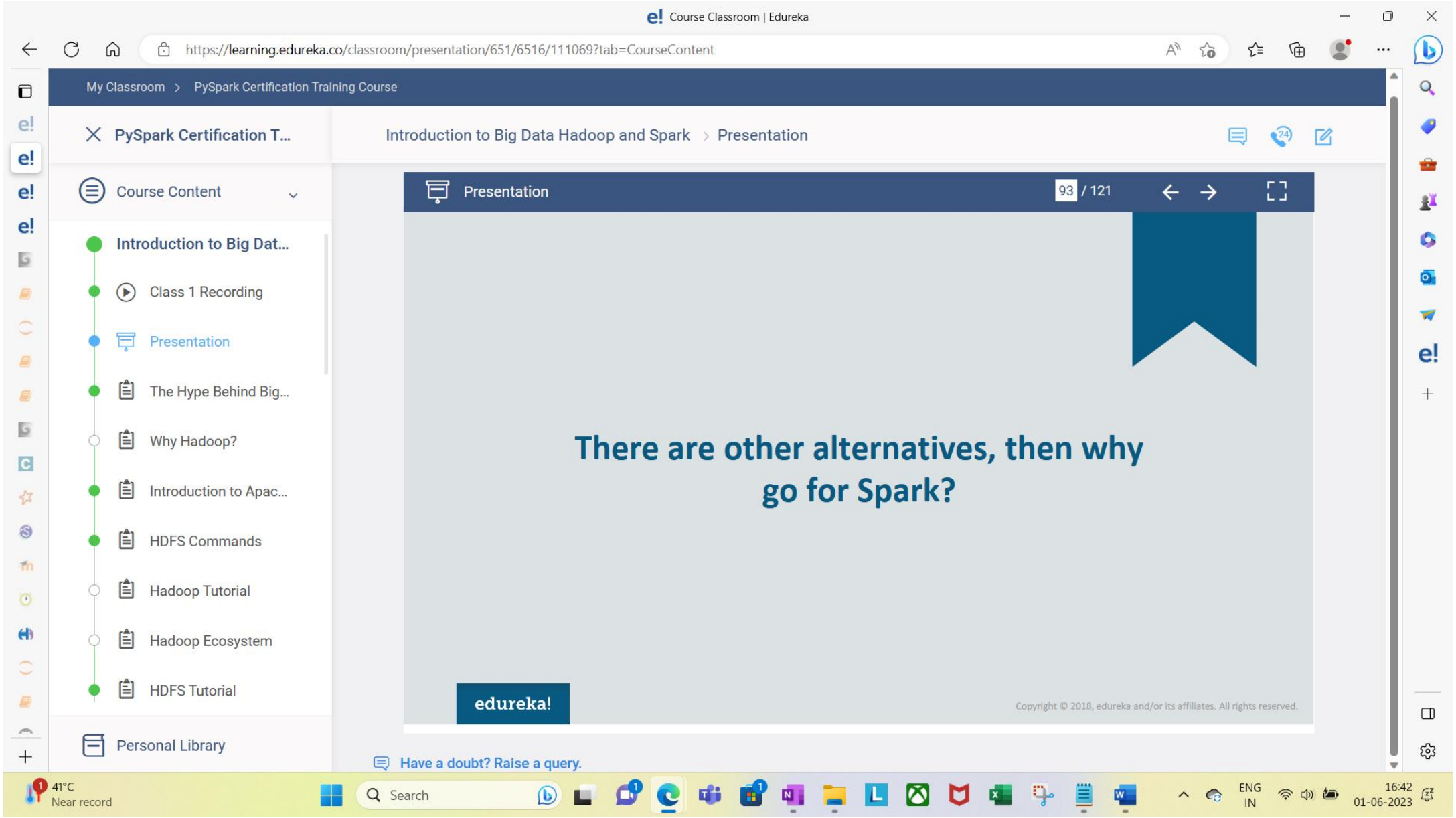
Presentation 92 / 121

Real-Time Processing



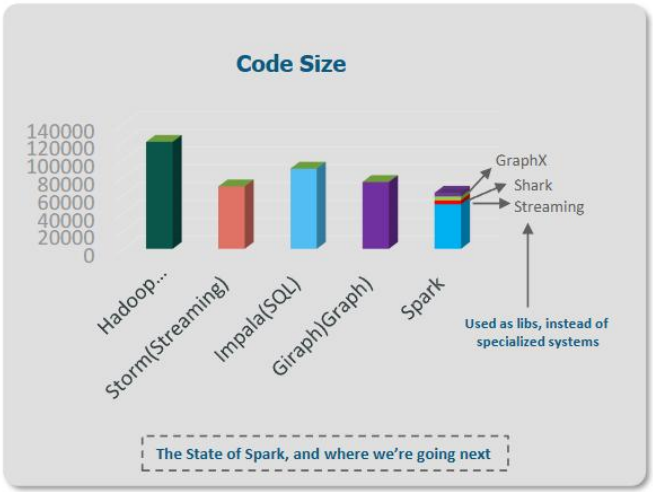
edureka!

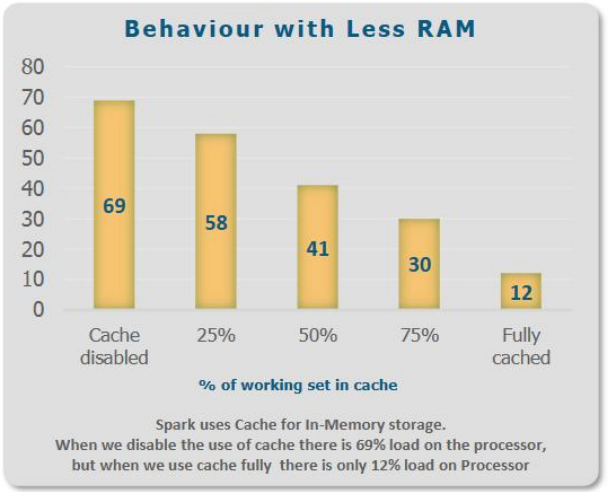
Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
93 / 121
There are other alternatives, then why go for Spark?
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

Presentation 94 / 121

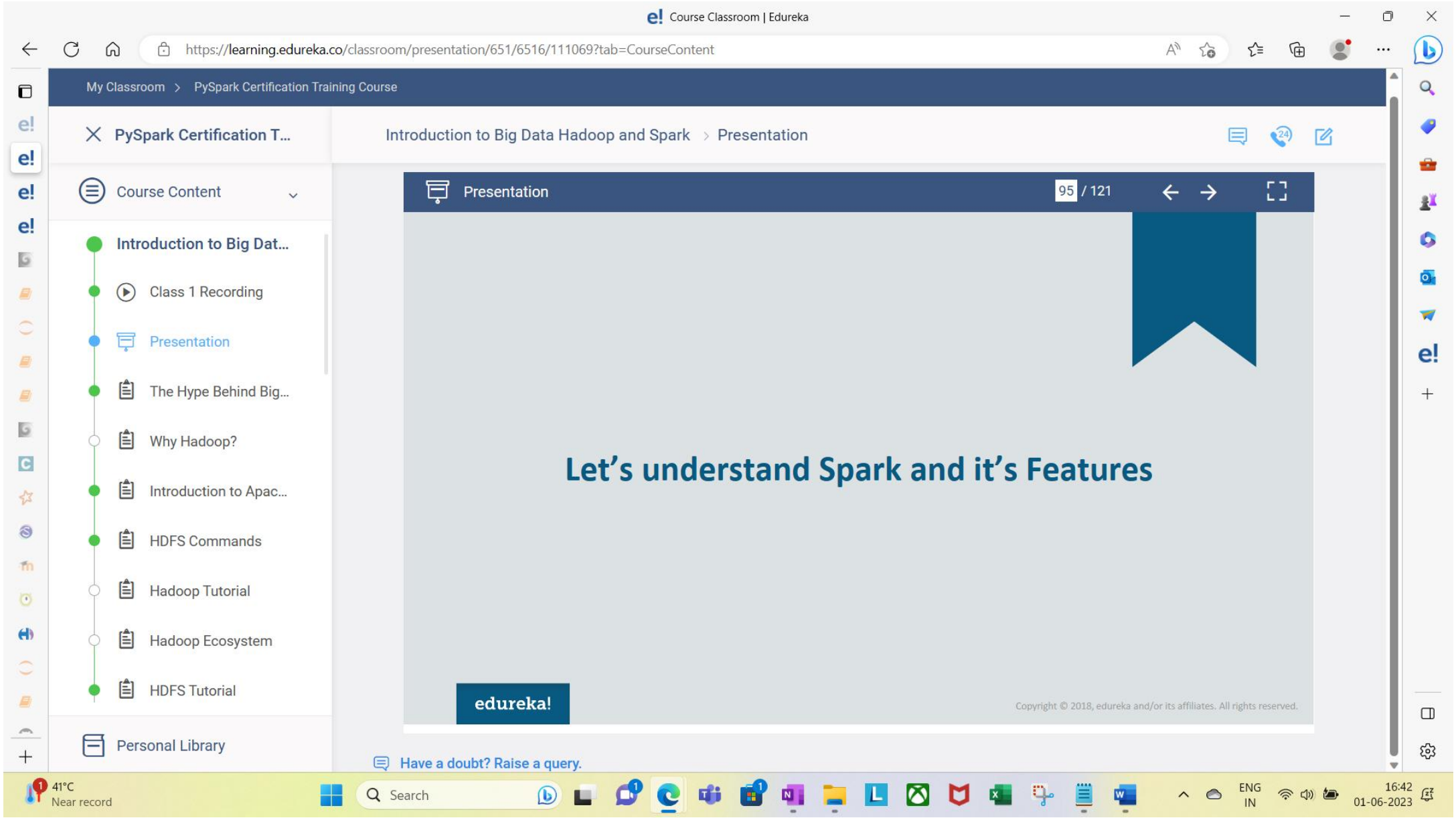
How Spark Differs?





edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
95 / 121
Let's understand Spark and it's Features
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:42
01-06-2023

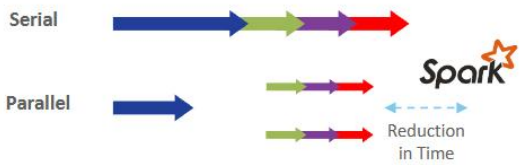
Presentation 96 / 121

What is Spark?

- Apache Spark is an *open-source cluster-computing framework* for both *Batch* as well as *real time processing*

Real Time Processing In Spark

- Spark provides an interface for *programming entire clusters with implicit data parallelism* and *fault-tolerance*



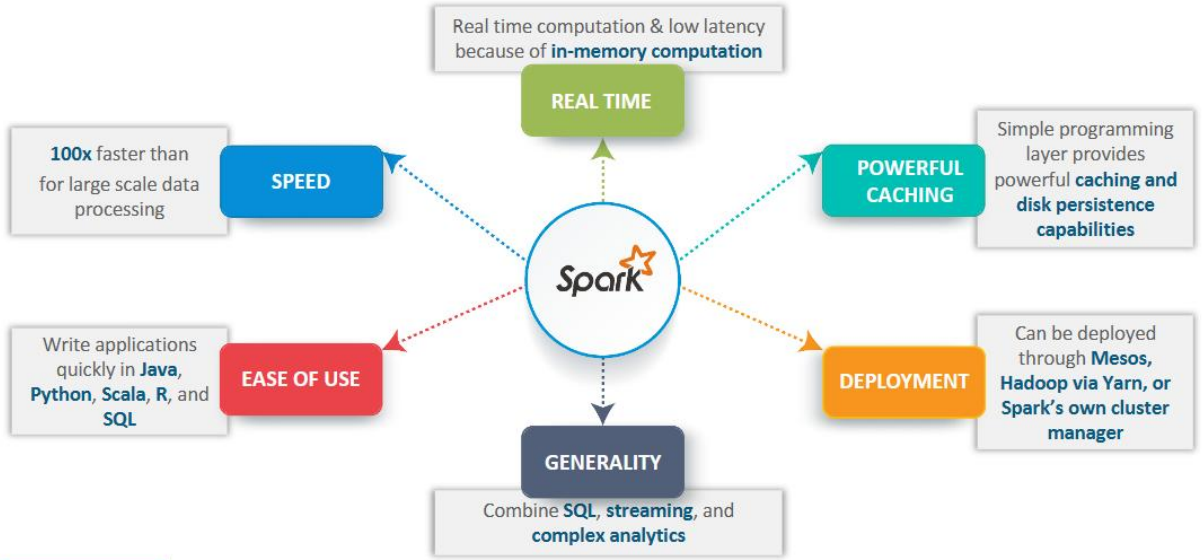
Data Parallelism in Spark

edureka!

Have a doubt? Raise a query.

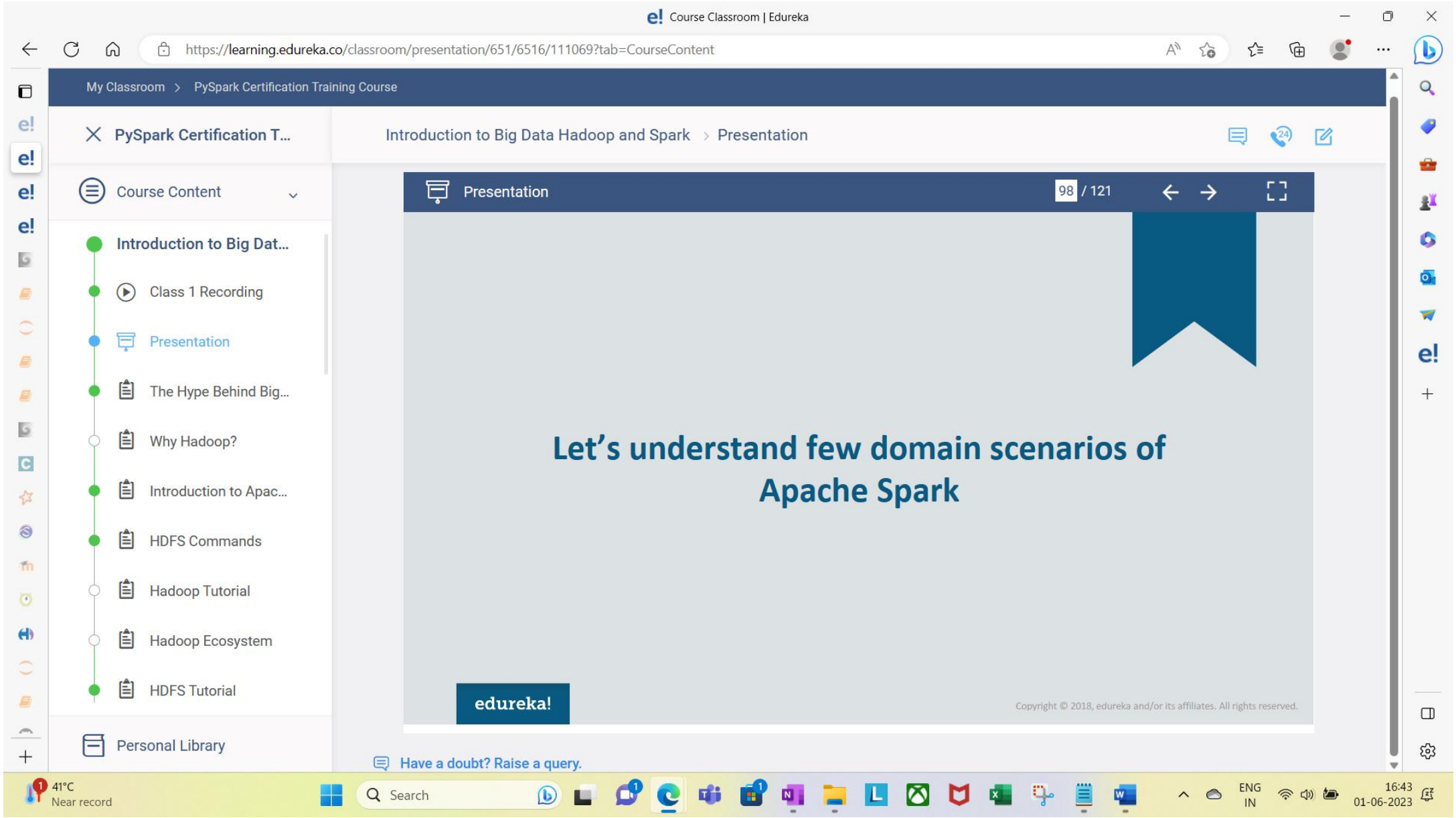
Presentation 97 / 121

Spark Features



edureka!

Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
98 / 121
Let's understand few domain scenarios of Apache Spark
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:43
01-06-2023

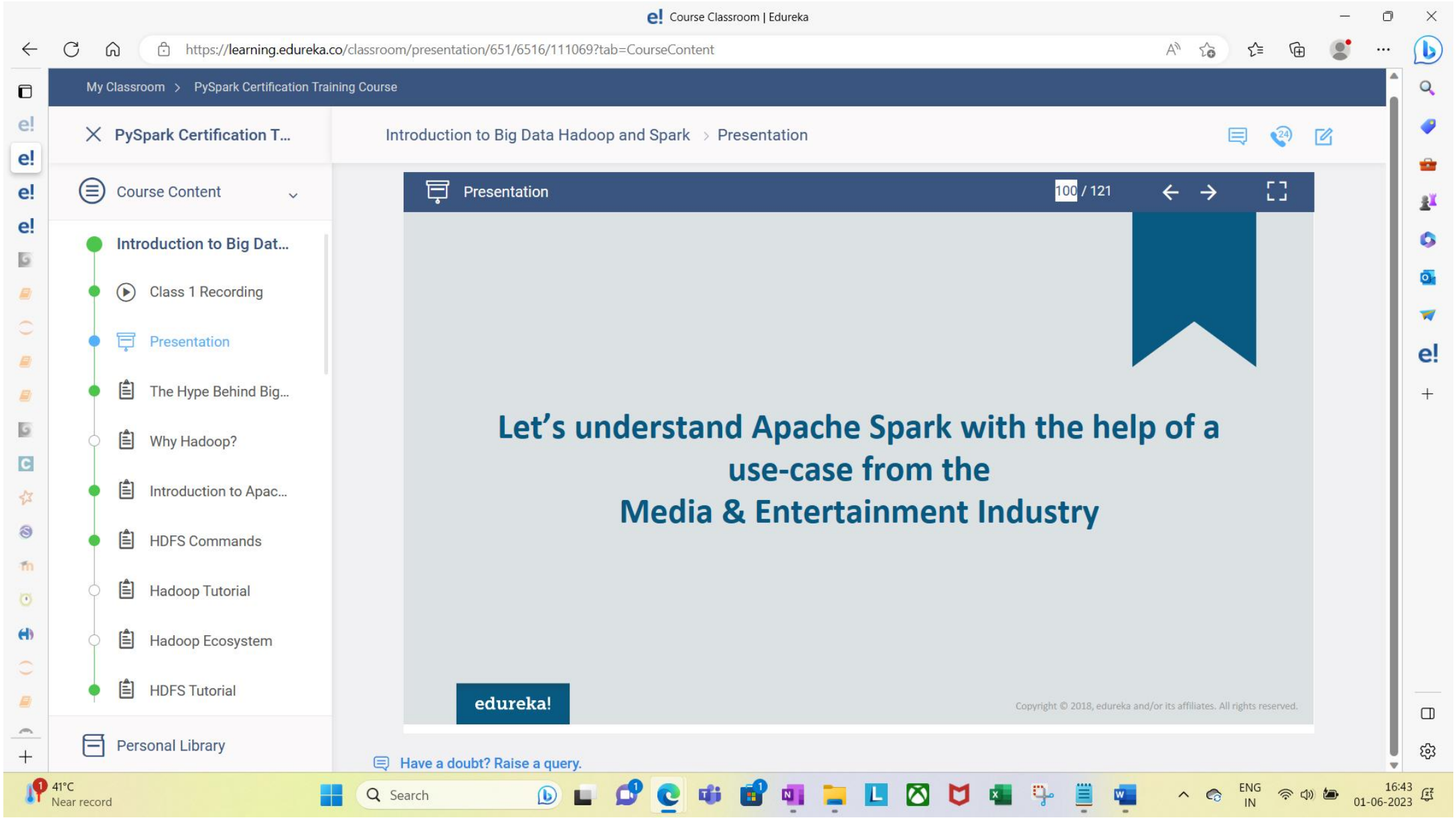
Presentation 99 / 121

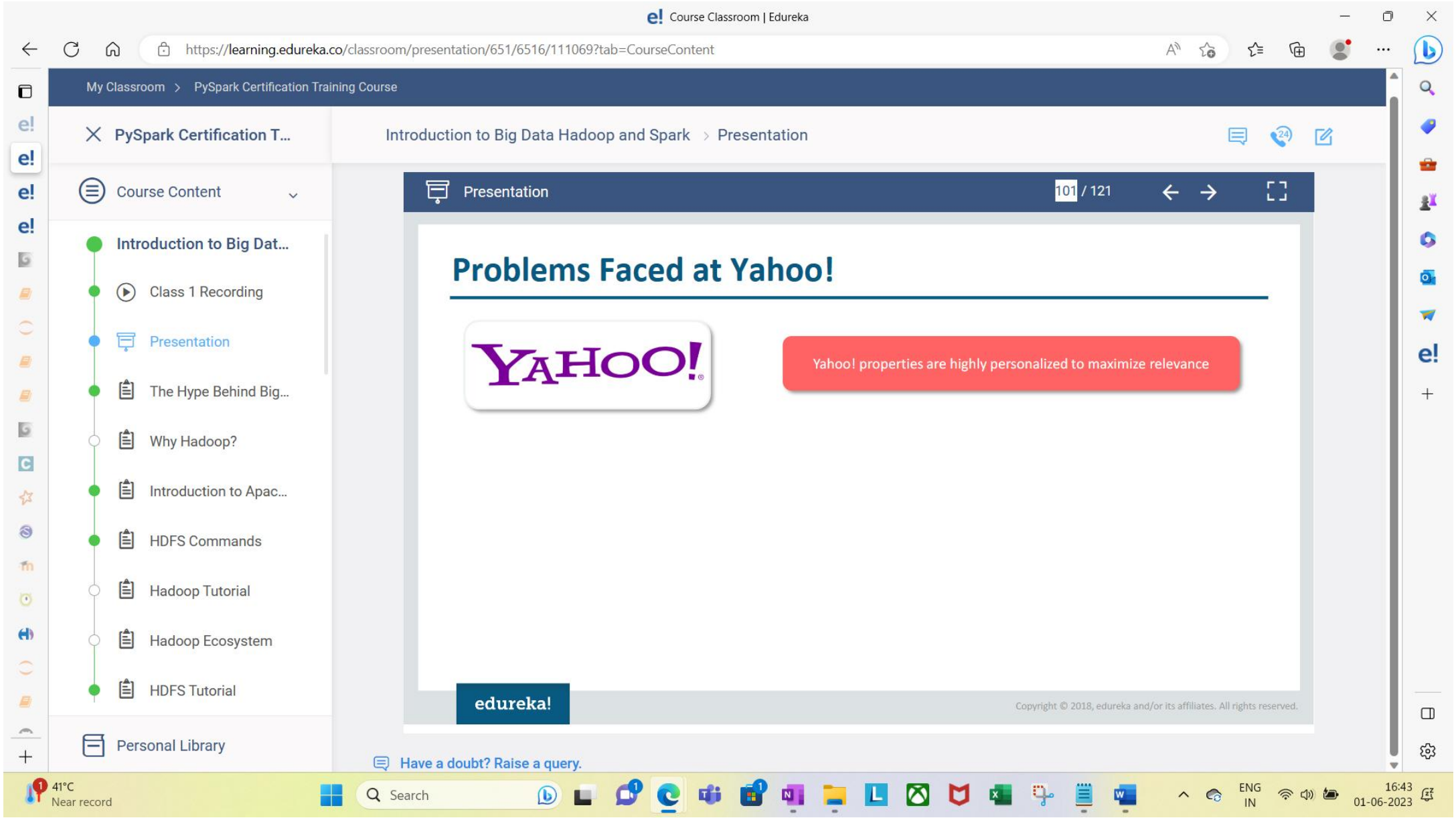
Apache Spark Domain Scenarios

- From small startups to *Fortune 500s* all are adopting Apache Spark to *build, scale* and *innovate* their big data applications

edureka!

Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
100 / 121
Let's understand Apache Spark with the help of a use-case from the Media & Entertainment Industry
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:43
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
101 / 121
Problems Faced at Yahoo!
YAHOO!
Yahoo! properties are highly personalized to maximize relevance
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Near record
Search
ENG
IN
16:43
01-06-2023

Presentation 102 / 121

Problems Faced at Yahoo!

YAHOO!

Yahoo! properties are highly personalized to maximize relevance

1. The *algorithms* used to provide personalization(targeted advertisement and personalized content) are ***highly sophisticated***

edureka!

Have a doubt? Raise a query.

Presentation 103 / 121

Problems Faced at Yahoo!

YAHOO!

Yahoo! properties are highly personalized to maximize relevance

1. The *algorithms* used to provide personalization(targeted advertisement and personalized content) are ***highly sophisticated***
2. *Relevance Model* must be *updated* frequently as stories, etc with change in time

edureka!

Have a doubt? Raise a query.

Presentation 104 / 121

Problems Faced at Yahoo!

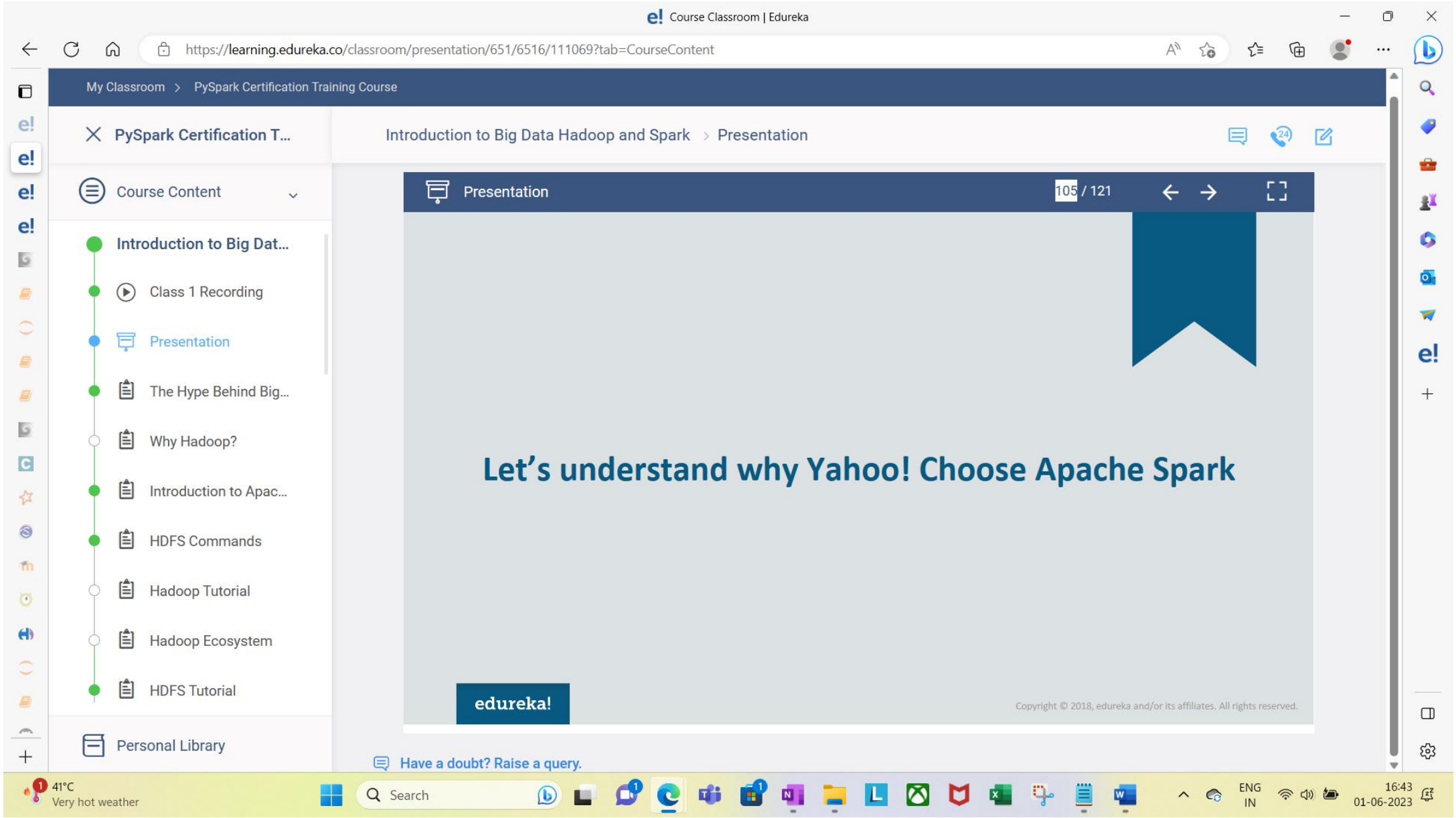
YAHOO!

Yahoo! properties are highly personalized to maximize relevance

1. The *algorithms* used to provide personalization(targeted advertisement and personalized content) are ***highly sophisticated***
2. *Relevance Model* must be *updated* frequently as stories, etc with change in time
3. Yahoo! has over *150 petabytes* of data stored on a *35,000-node Hadoop cluster* which should be accessed efficiently to:
   - *Avoid latency* caused by data movement and
   - *Gain insights* from data in cost-effective manner

edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
105 / 121
Let's understand why Yahoo! Choose Apache Spark
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Very hot weather
Search
ENG
IN
16:43
01-06-2023

Presentation 106 / 121

Why Apache Spark?

Yahoo! looked to Spark to improve ***performance*** of its ***iterative model training***

edureka!

Have a doubt? Raise a query.

Presentation 107 / 121

Why Apache Spark?

Yahoo! looked to Spark to improve ***performance*** of its ***iterative model training***

The machine learning algorithm for news personalization required *15000 lines of C++ code*

edureka!

Have a doubt? Raise a query.

Presentation 108 / 121

Why Apache Spark?

Yahoo! looked to Spark to improve ***performance*** of its ***iterative model training***

The machine learning algorithm for news personalization required *15000 lines of C++ code*

PySpark has reduced the length of code to more than half the size

edureka!

Have a doubt? Raise a query.

Why Apache Spark?

Yahoo! looked to Spark to improve ***performance*** of its ***iterative model training***

PySpark has reduced the length of code to more then half the size

The algorithm was ready for production use in just *30 minutes* of training, on a *hundred million datasets*

Its rich API is available in several programming languages, has resilient in-memory storage options and is compatible with Hadoop through YARN and the Spark-YARN project

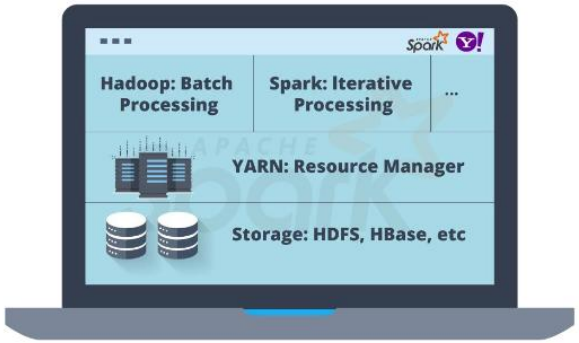
edureka!

Have a doubt? Raise a query.

Presentation 110 / 121

Spark at Yahoo!

- It uses Apache Spark for *personalizing* its *news webpages* and for *targeted advertising*
- It uses *machine learning algorithms* that run on Apache Spark
  - to find out what kind of news - users are interested to read
  - For categorizing the news stories to find out what kind of users would be interested in reading each category of news
- Spark runs over *Hadoop YARN* to use existing data & clusters



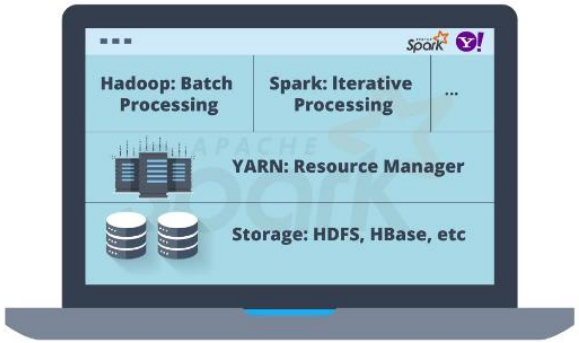
edureka!

Have a doubt? Raise a query.

Presentation 111 / 121

Spark at Yahoo!

- *Spark's API* & *MLlib ease* development of ML algorithms
- Spark *reduces latency* of model training via in-memory *RDD*
- *YARN* brings *Hadoop datasets* & *clusters* at scientist's fingertip



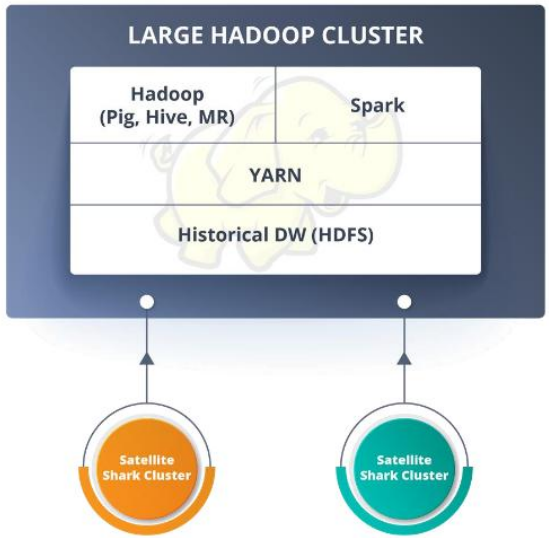
edureka!

Have a doubt? Raise a query.

Presentation 112 / 121

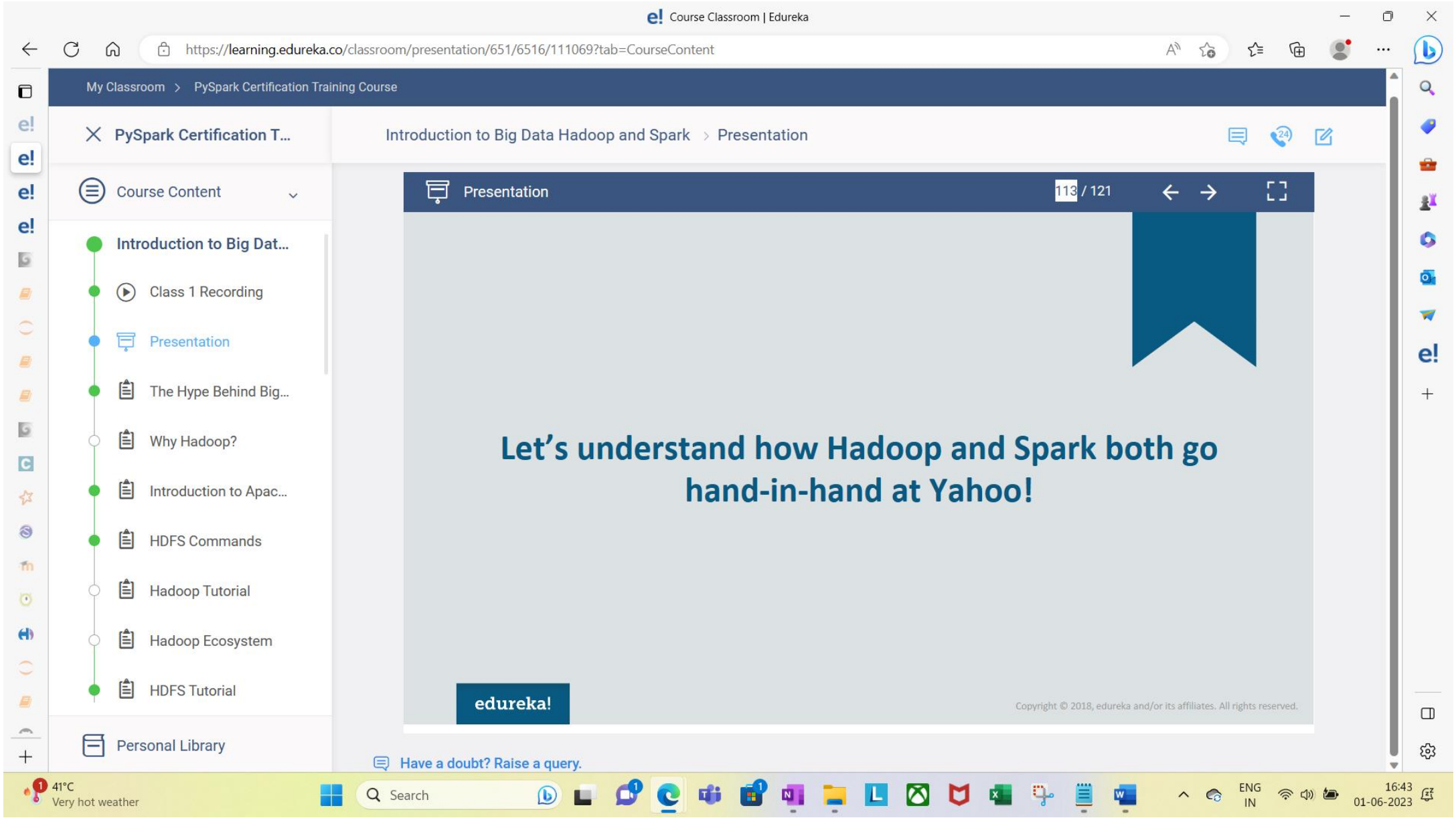
Spark at Yahoo! (Ad Analytics)

- Yahoo! Ads wanted *interactive BI* on *terabytes of data*
- Chose Shark (Hive on Spark) to provide this through standard *Hive server API + Tableau*
- **Result:** Were able to *query* their *ad visit data interactively*



edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
113 / 121
Let's understand how Hadoop and Spark both go hand-in-hand at Yahoo!
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Very hot weather
Search
ENG
IN
16:43
01-06-2023

Presentation 114 / 121

Hadoop + Spark at Yahoo!

Yahoo! tested Spark's performance relative to Hadoop using an e-commerce pilot project

edureka!

Have a doubt? Raise a query.

Presentation 115 / 121

Hadoop + Spark at Yahoo!

Yahoo! tested Spark's performance relative to Hadoop using an e-commerce pilot project

Apache Spark

The pilot was *successfully implemented* with *30 lines of code* (*Spark's API*) and executed in *14 minutes* on just *10 servers*

hadoop

The equivalent Hadoop implementation took *106 minutes*

edureka!

Have a doubt? Raise a query.

Presentation 116 / 121

Hadoop + Spark at Yahoo!

Yahoo! tested Spark's performance relative to Hadoop using an e-commerce pilot project

The pilot was *successfully implemented* with *30 lines of code* (*Spark's API*) and executed in *14 minutes* on just *10 servers*

The equivalent Hadoop implementation took *106 minutes*

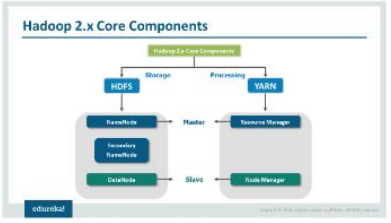
- While these improvements are impressive, Yahoo! isn't abandoning its Hadoop cluster for Spark
- There is a clear *need for both* types of workloads
- *Spark* will be the *preferred* technology for *iterative processing*, while *Hadoop* continues to fulfil its niche for *batch data processing* tasks
- What's interesting is that both types of tasks run on the same Hadoop cluster through YARN

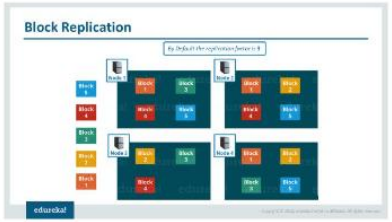
edureka!

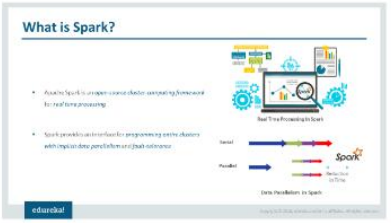
Have a doubt? Raise a query.

Presentation 117 / 121

Summary

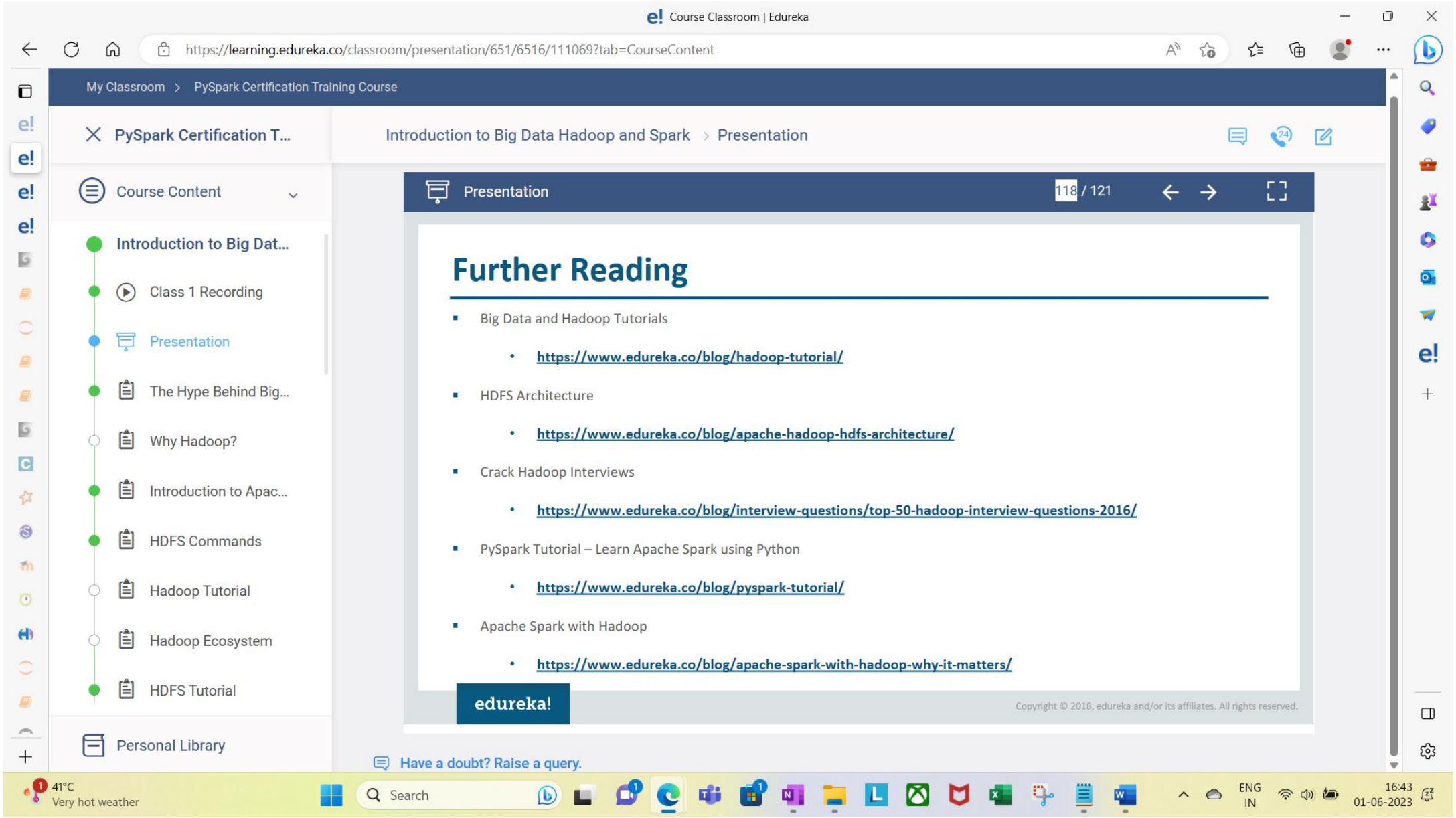






edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6516/111069?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Introduction to Big Data Hadoop and Spark > Presentation
Course Content
Introduction to Big Dat...
Class 1 Recording
Presentation
The Hype Behind Big...
Why Hadoop?
Introduction to Apac...
HDFS Commands
Hadoop Tutorial
Hadoop Ecosystem
HDFS Tutorial
Personal Library
Presentation
118 / 121
Further Reading
Big Data and Hadoop Tutorials
https://www.edureka.co/blog/hadoop-tutorial/
HDFS Architecture
https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/
Crack Hadoop Interviews
https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/
PySpark Tutorial – Learn Apache Spark using Python
https://www.edureka.co/blog/pyspark-tutorial/
Apache Spark with Hadoop
https://www.edureka.co/blog/apache-spark-with-hadoop-why-it-matters/
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
41°C
Very hot weather
Search
ENG
IN
16:43
01-06-2023