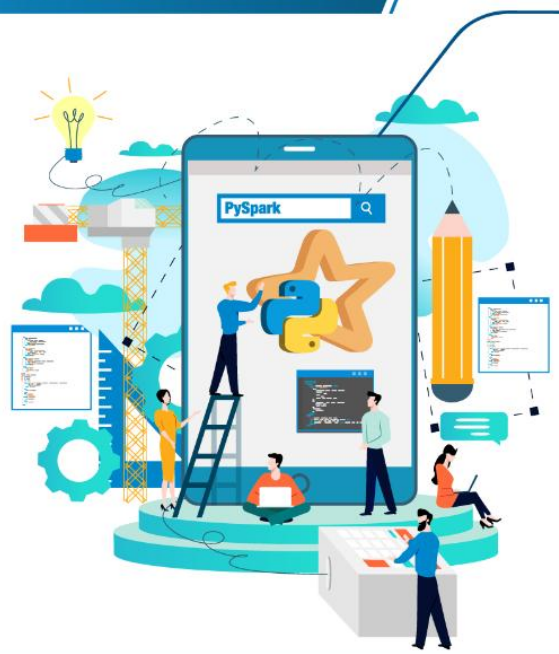
Presentation 3 / 81

# COURSE OUTLINE




## MODULE 04

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 / 81

# Objectives

After completing this module, you should be able to:

- Spark Components and It's Architecture
- Spark Deployment Modes
- Introduction to PySpark Shell
- Submitting PySpark Job
- Spark Web UI
- Writing your first PySpark Job Using Jupyter Notebook
- Data Ingestion using Sqoop


edureka!

Presentation 5 / 81

## 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!



We have already discussed in Module 1 about the problems faced at Yahoo for Dynamic Content and faster availability of Customized data.

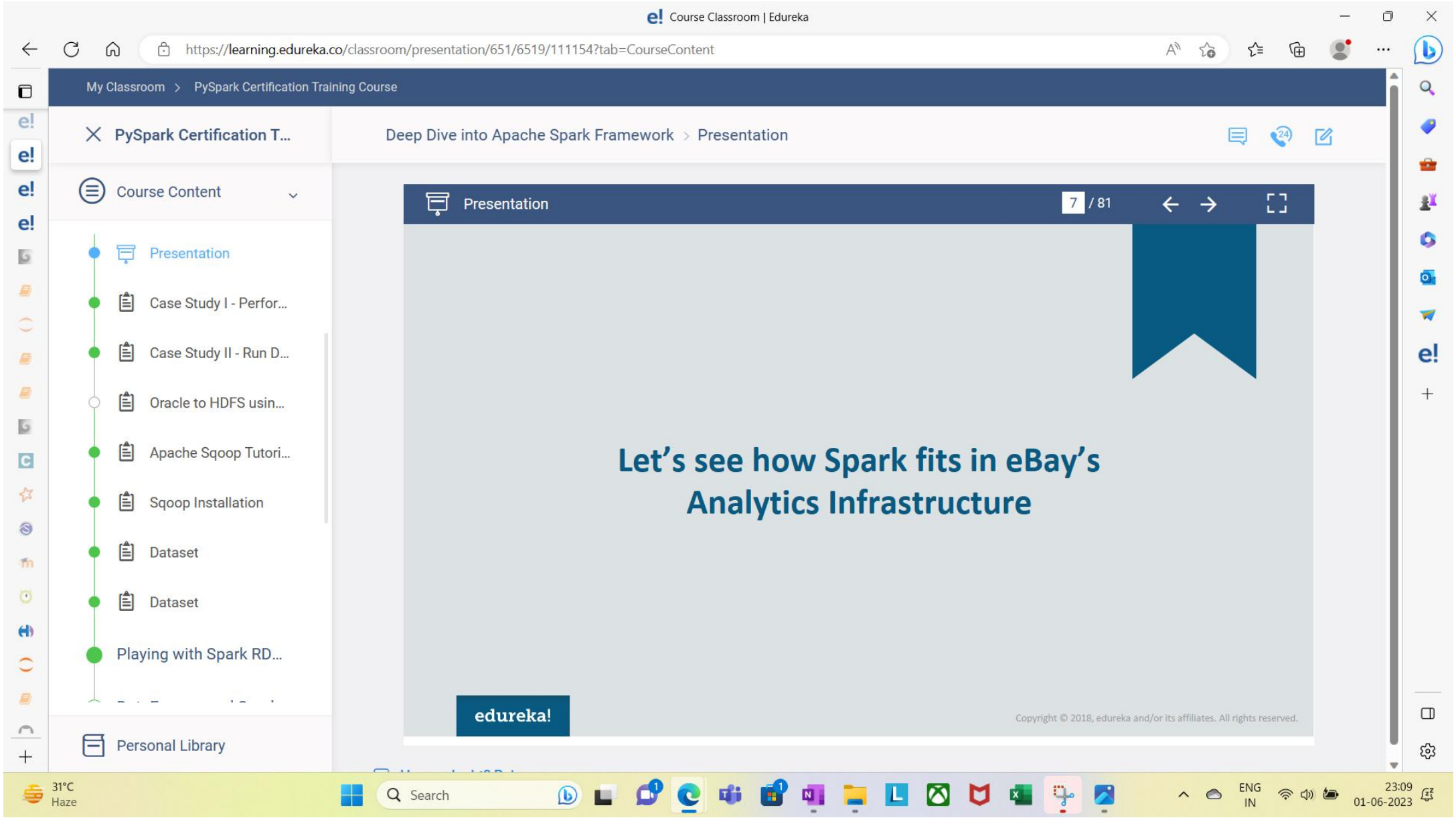
Presentation 6 / 81

- The problem domain includes
  - ***Operational management*** (performance, errors, anomaly detection)
  - ***Triaging*** (Root Cause Analysis) and
  - ***Business monitoring*** (customer behaviour, click stream analytics)
- How much data did eBay collected?
  - In 2009, it was around 10 TB/day, now more than ***500 TB/day***

**Note**: All these problems were arising mainly due to ***late processing of data***

*Why are my analysis not helping the company? We need data and results in a fast and efficient way*

Now, let us have a look at ebay's Analytics Infrastructure and decide how Spark can help us achieve needed efficiency within optimal time frame


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
7 / 81
Let's see how Spark fits in eBay's Analytics Infrastructure
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:09 01-06-2023

Presentation 8 / 81

## How Big is eBay?


edureka!

# Problems in eBay's Infrastructure

- The problem domain includes
  - ***Operational management*** (performance, errors, anomaly detection)
  - ***Triaging*** (Root Cause Analysis) and
  - ***Business monitoring*** (customer behaviour, click stream analytics)
- How much data did eBay collected?
  - In 2009, it was around 10 TB/day, now more than ***500 TB/day***

All these problems were arising mainly due to ***late processing of data***

*Why are my analysis not helping the company? We need data and results in a fast and efficient way*

edureka!

Presentation 10 / 81

# Apache Spark at eBay

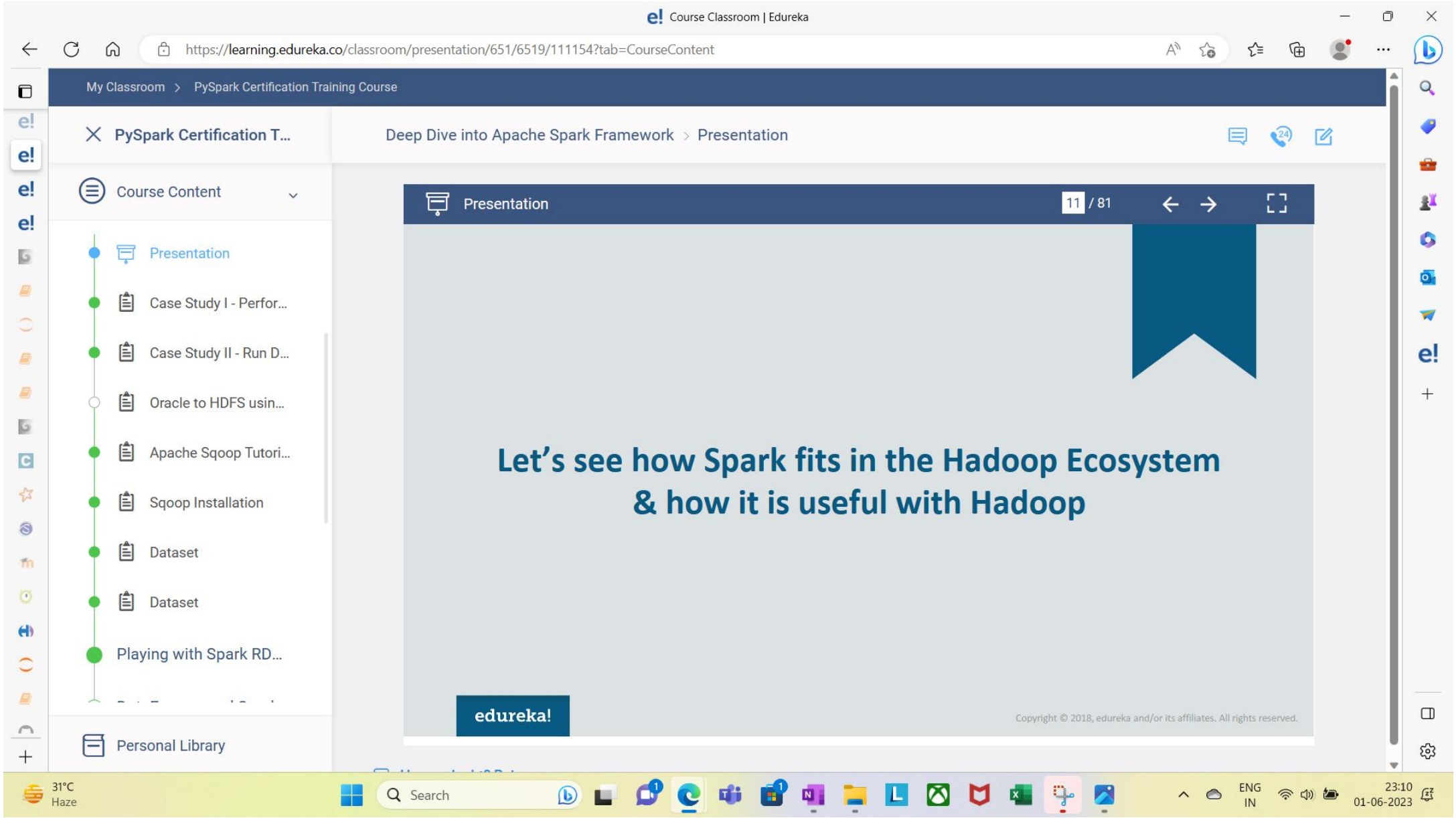
With **Hadoop** as base storage and **Spark** for Processing they were able to convert 24-48 hr ETL process to **near real time**

eBay's analysts are able to leverage clusters approaching the range **of 2000 nodes, 100TB of RAM,** and **20,000 cores**


**Near real time analysis** based on **Kafka** (messaging), **Spark Streaming** (stream processing) and **Spark SQL** (data-preparation)

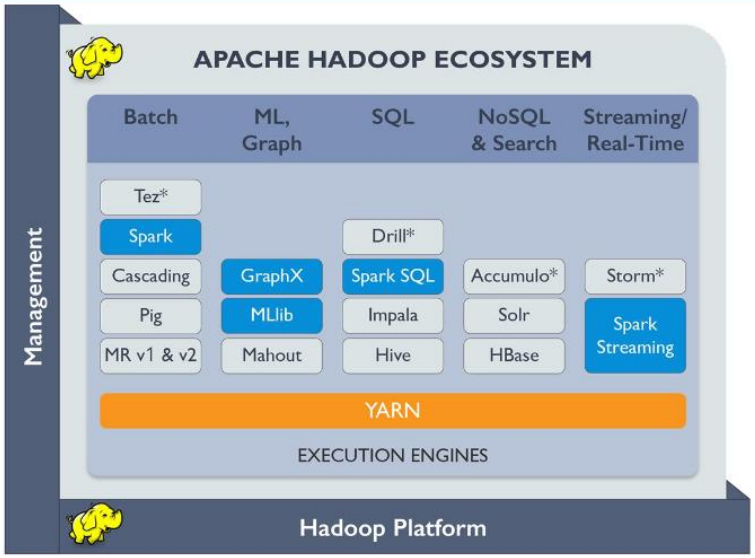
eBay's developers are writing Multiple Domain codes helping through complex **data modelling** and **data scoring**

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
11 / 81
Let's see how Spark fits in the Hadoop Ecosystem
& how it is useful with Hadoop
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:10
01-06-2023

Presentation 12 / 81

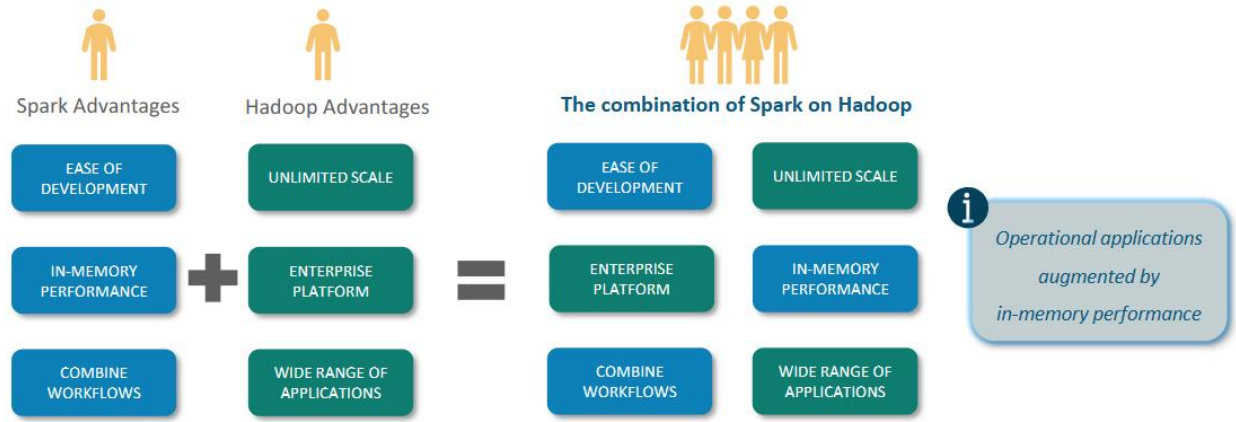
# Spark in Hadoop Ecosystem



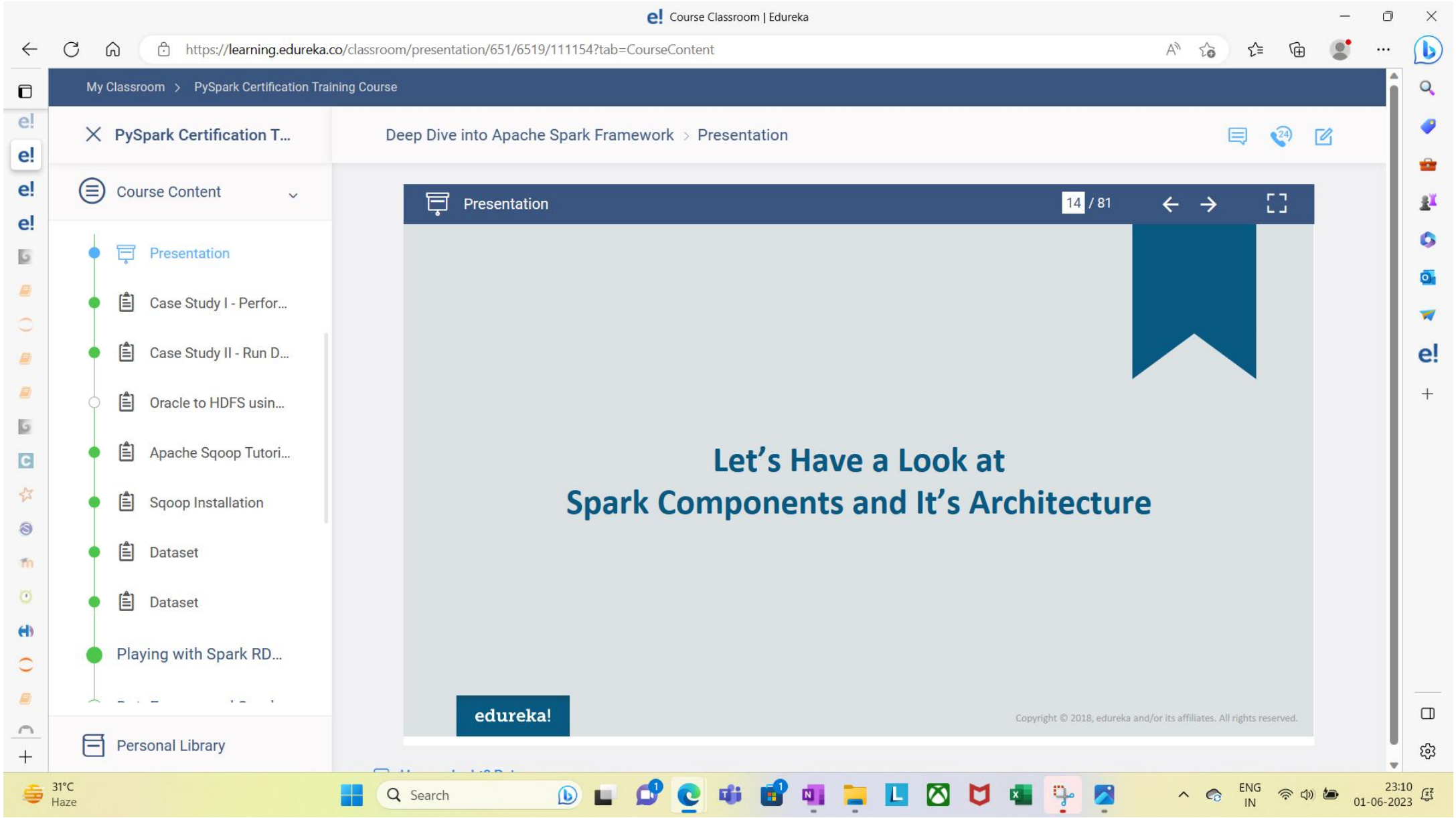

edureka!

Presentation 13 / 81

## Spark and Hadoop Together

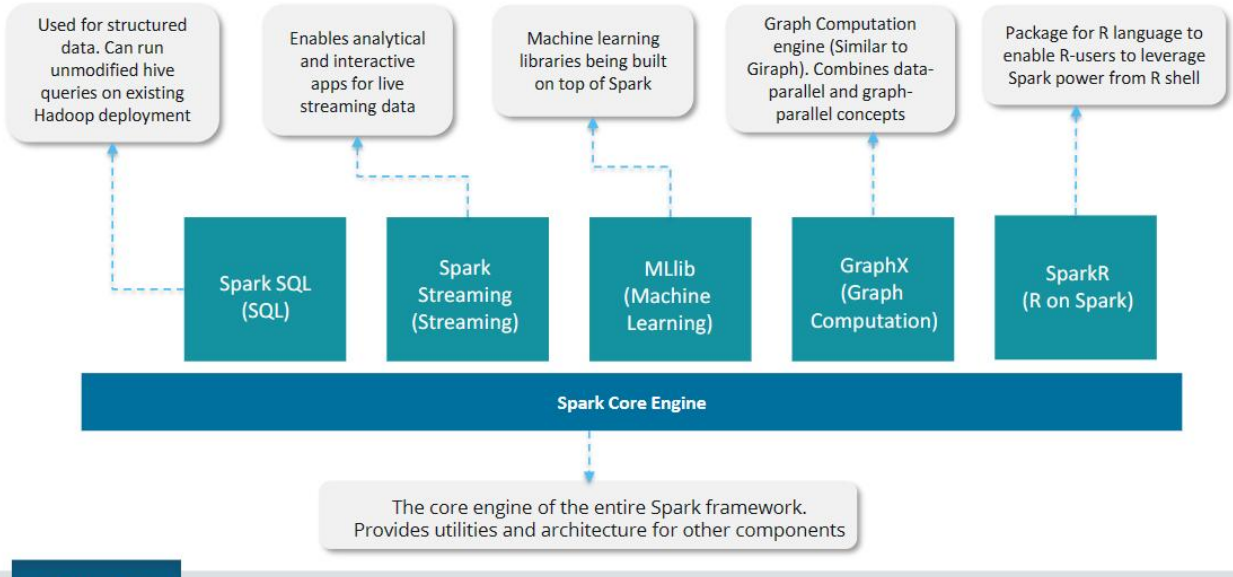



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
14 / 81
Let's Have a Look at
Spark Components and It's Architecture
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:10
01-06-2023

Presentation 15 / 81

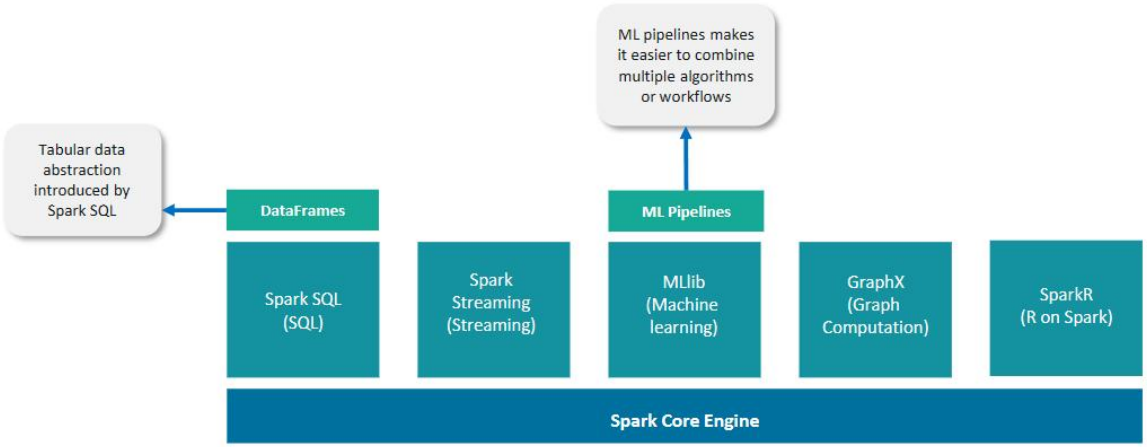
# Spark Components




edureka!

Presentation 16 / 81

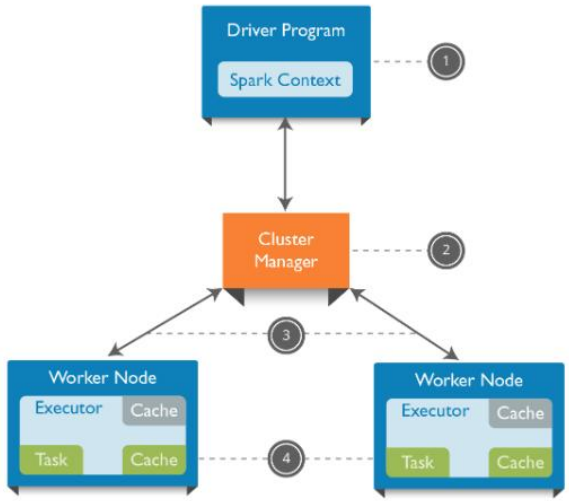
# Spark Components




edureka!

Presentation 17 / 81

# Spark Architecture and SparkContext

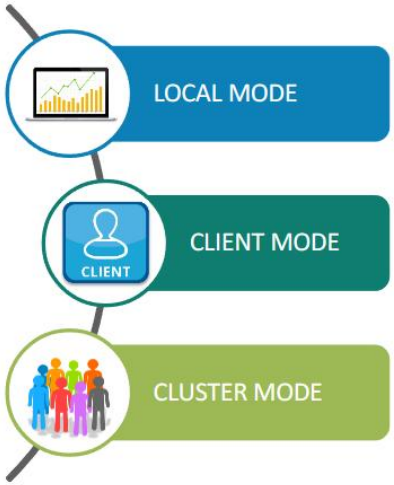



1. ***SparkContext*** object in *driver program coordinates all the distributed process* and *allows resource allocation*
2. Cluster Managers provide Executors = JVM process with logic
3. SparkContext object *sends the application to executors*
4. SparkContext *executes tasks in each executor*

edureka!

Presentation 18 / 81

# Spark Modes

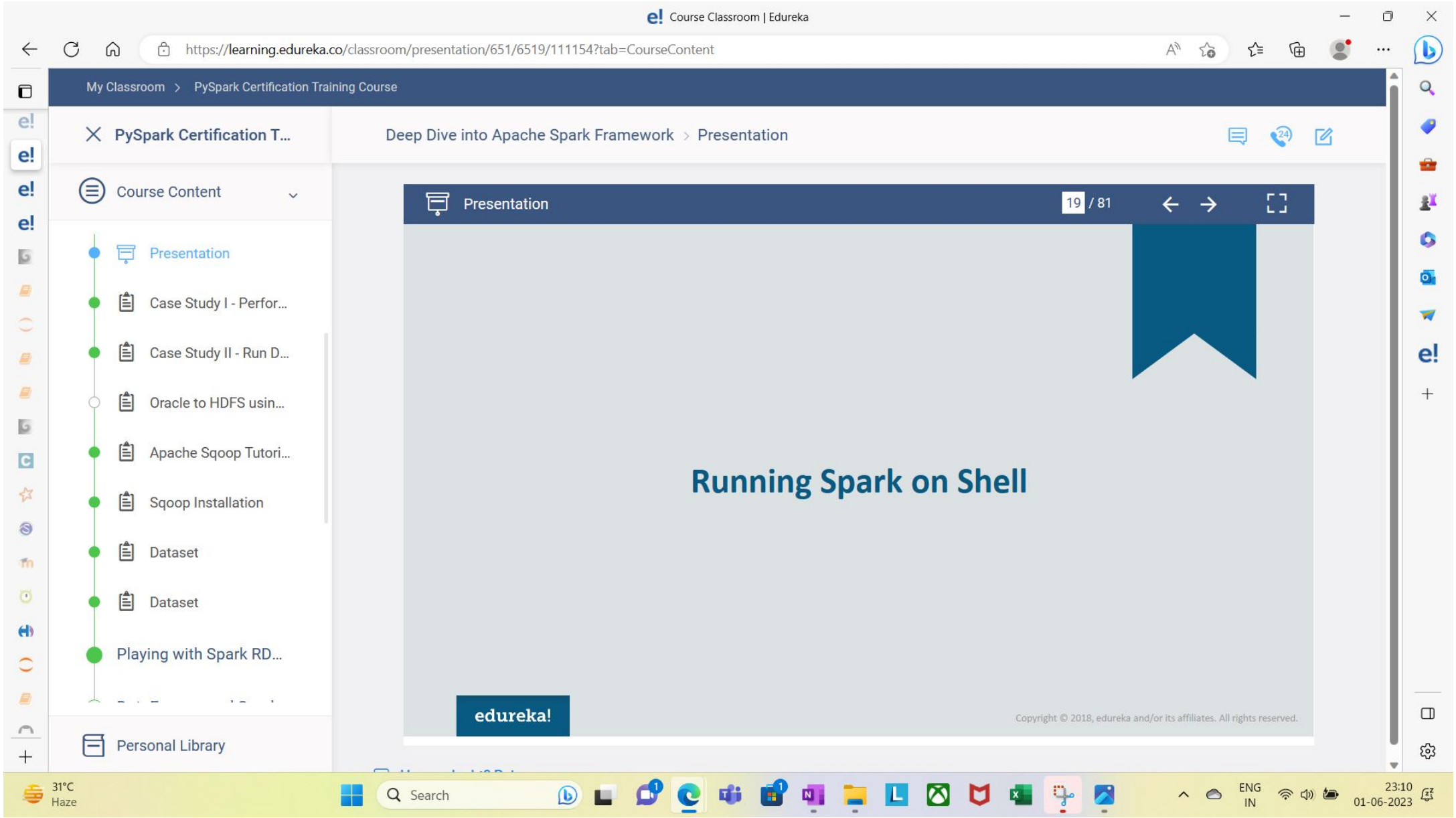



/spark-submit --class org.apache.spark.examples.SparkPi --**master local** examples/target/python-2.7.5/spark-examples-1.5.2-hadoop2.2.0.jar

/spark-submit --class org.apache.spark.examples.SparkPi --master spark://localhost.localdomain:7077 --**deploy-mode client** --executor-memory 1g --executor-cores 1 examples/target/ python-2.7.5 /spark-examples-1.5.2-hadoop2.2.0.jar

/spark-submit --class org.apache.spark.examples.SparkPi --master spark://localhost.localdomain:7077 --**deploy-mode cluster** --executor-memory 1g --executor-cores 1 examples/target/ python-2.7.5 /spark-examples-1.5.2-hadoop2.2.0.jar

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
19 / 81
Running Spark on Shell
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:10
01-06-2023

Presentation 20 / 81

# Spark Shell : Python

```
[edureka_294428@ip-20-0-41-202 ~]$ pyspark2
Python 2.7.5 (default, Aug  4 2017, 00:39:18)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-16)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "ERROR".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.1.0.cloudera2
      /_/

Using Python version 2.7.5 (default, Aug  4 2017 00:39:18)
SparkSession available as 'spark'.
```

edureka!

Presentation 21 / 81

# Running Spark Programs on Terminal

Type “vi filename.py” to create a Python file

```
[edureka_294428@ip-20-0-41-202 ~]$ vi DemoPy.py
```

edureka!

Presentation 22 / 81

## Running Spark Programs on Terminal

Type “vi filename.py” to create a Python file

```
[edureka_294428@ip-20-0-41-202 ~]$ vi DemoPy.py
```

- Write the code
- Press Esc, then type “:wq” to come out of the file back to Terminal

```
print("Hello World")
print("Welcome to edureka")
print("Happy Learning")
```

edureka!

Presentation 23 / 81

# Running Spark Programs on Terminal

Type "vi filename.py" to create a Python file

```
[edureka_294428@ip-20-0-41-202 ~]$ vi DemoPy.py
```

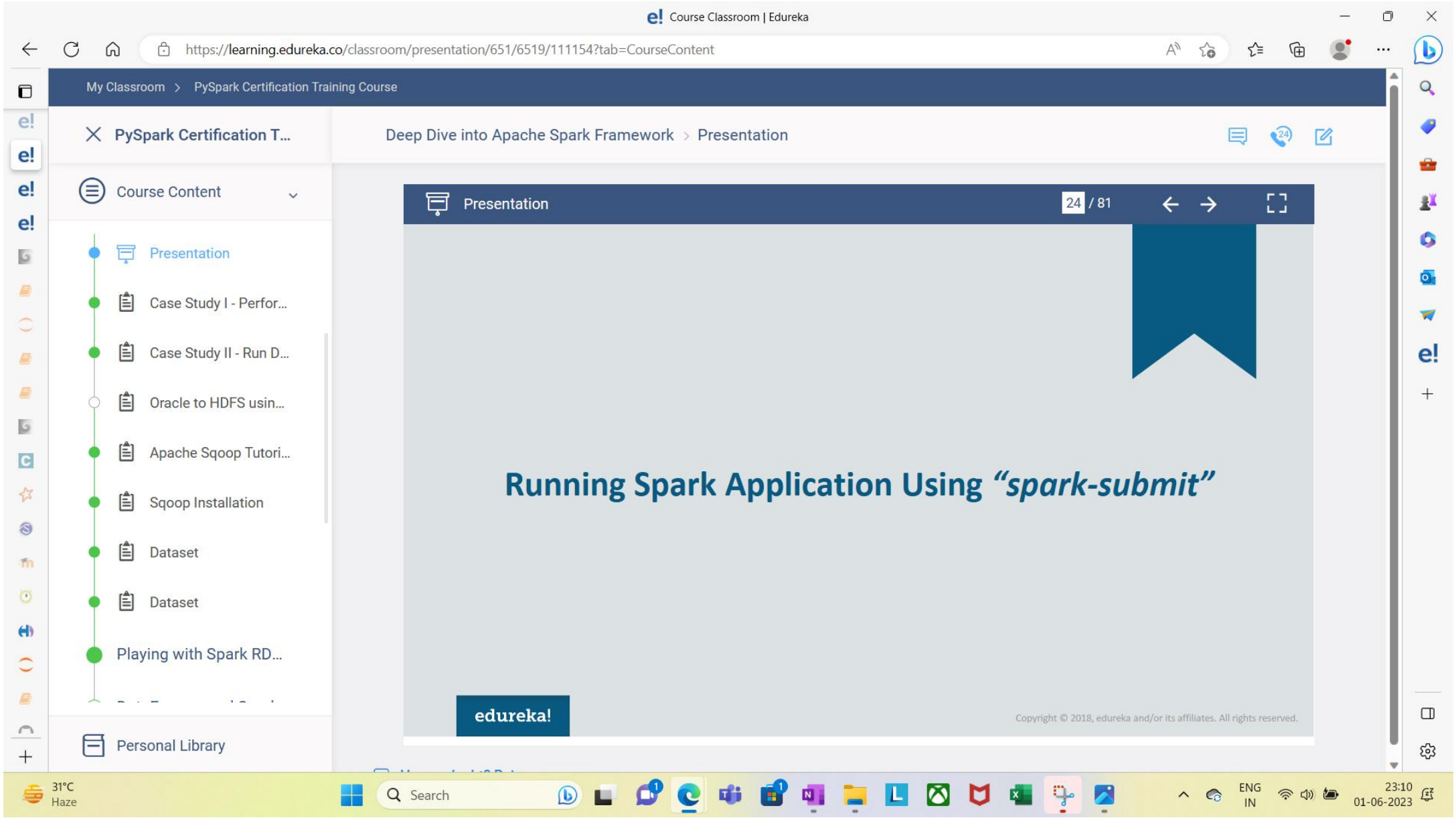
- Write the code
- Press Esc, then type ":wq" to come out of the file back to Terminal

```
print("Hello World")
print("Welcome to edureka")
print("Happy Learning")
```

To Execute : "python filename.py"

```
[edureka_294428@ip-20-0-41-202 ~]$ python DemoPy.py
Hello World
Welcome to edureka
Happy Learning
[edureka_294428@ip-20-0-41-202 ~]$
```

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
24 / 81
Running Spark Application Using "spark-submit"
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:10
01-06-2023

Presentation 25 / 81

# Writing wordcount.py Program

```
import sys

from pyspark import SparkContext, SparkConf

if __name__ == "__main__":

  # create Spark context with Spark configuration
  conf = SparkConf().setAppName("Word Count - Python").set("spark.hadoop.yarn.resourcemanager.address", "192.168.0.104:8032")
  sc = SparkContext(conf=conf)

  # read in text file and split each document into words
  words = sc.textFile("/user/edureka_321047/Practice/PySpark.xxx").flatMap(lambda line: line.split(" "))

  # count the occurrence of each word
  wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a +b)

  wordCounts.saveAsTextFile("/user/edureka_321047/Practice/output")
```

Provide the HDFS path for the file on which you want to perform the read operation

edureka!

Presentation 26 / 81

# Writing wordcount.py Program

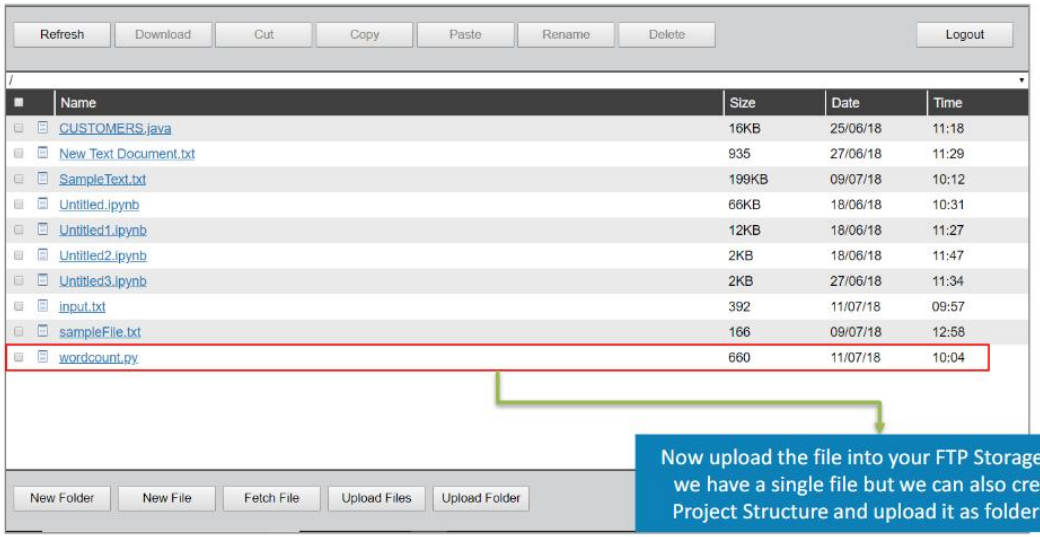
```
import sys

from pyspark import SparkContext, SparkConf

if __name__ == "__main__":

  # create Spark context with Spark configuration
  conf = SparkConf().setAppName("Word Count - Python").set("spark.hadoop.yarn.resourcemanager.address", "192.168.0.104:8032")
  sc = SparkContext(conf=conf)

  # read in text file and split each document into words
  words = sc.textFile("/user/edureka_321047/Practice/PySpark.txt").flatMap(lambda line: line.split(" "))

  # count the occurrence of each word
  wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a +b)

  wordCounts.saveAsTextFile("/user/edureka_321047/Practice/output")
```

Provide the HDFS path for the output folder inside which you want the result to be saved

edureka!

Presentation 27 / 81

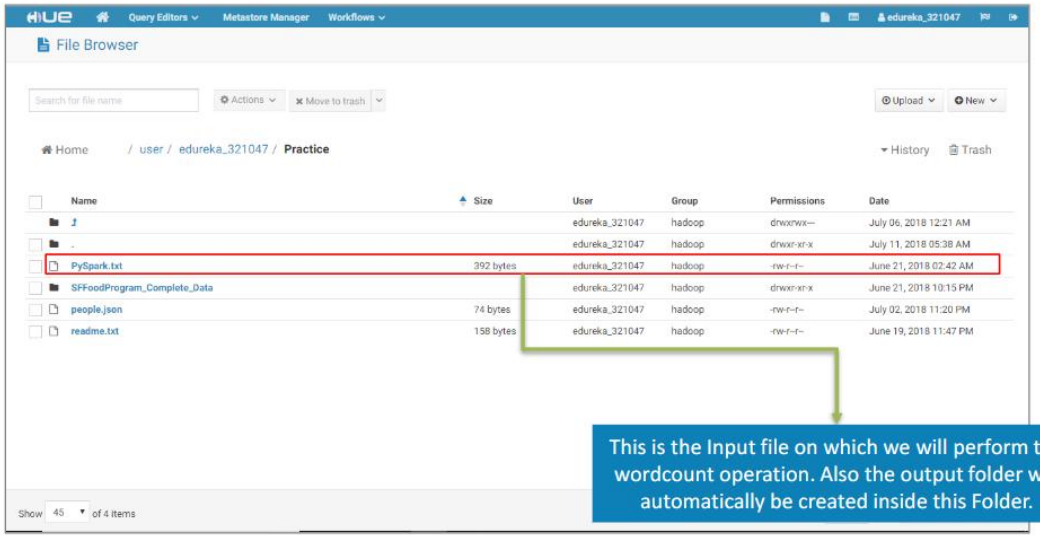
# Uploading wordcount.py On FTP




edureka!

Presentation 28 / 81

# Input File in HUE




edureka!

Presentation 29 / 81

# Running spark-submit On Python File

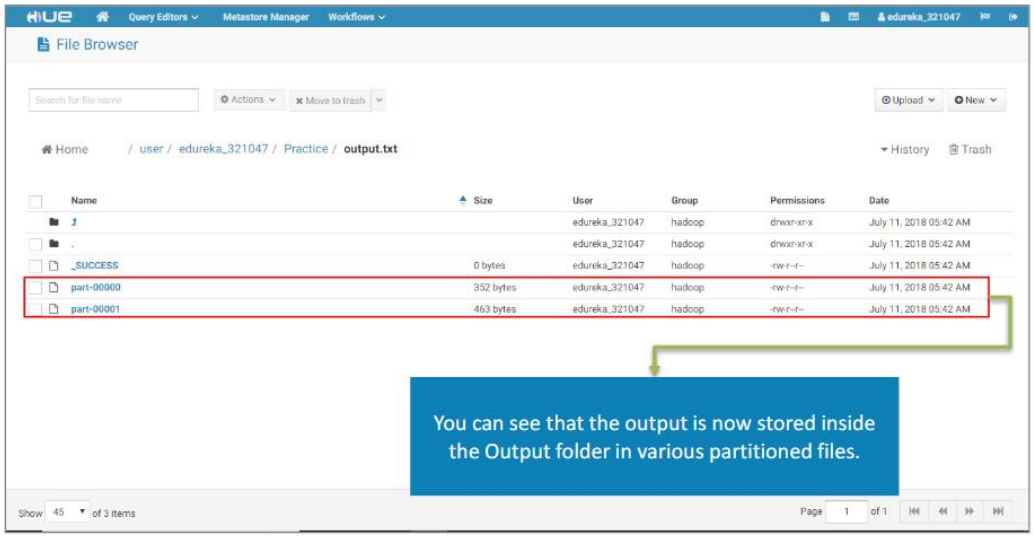
```
ip-20-0-41-190 login: edureka_321047
Password:
Last login: Wed Jul 11 06:03:20 on pts/64
[edureka_321047@ip-20-0-41-190 ~]$ spark2-submit wordcount.py
18/07/11 12:42:46 INFO spark.SparkContext: Running Spark version 2.1.0.cloudera2
18/07/11 12:42:47 INFO spark.SecurityManager: Changing view acls to: edureka_321047
18/07/11 12:42:47 INFO spark.SecurityManager: Changing modify acls to: edureka_321047
18/07/11 12:42:47 INFO spark.SecurityManager: Changing view acls groups to:
18/07/11 12:42:47 INFO spark.SecurityManager: Changing modify acls groups to:
18/07/11 12:42:47 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(edureka_321047); groups
 with view permissions: Set(); users  with modify permissions: Set(edureka_321047); groups with modify permissions: Set()
18/07/11 12:42:47 INFO util.Utils: Successfully started service 'sparkDriver' on port 44049.
18/07/11 12:42:47 INFO spark.SparkEnv: Registering MapOutputTracker
18/07/11 12:42:47 INFO spark.SparkEnv: Registering BlockManagerMaster
18/07/11 12:42:47 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
18/07/11 12:42:47 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/07/11 12:42:47 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-c2a963ca-b51f-40c1-9a26-1a8c02e5faf4
18/07/11 12:42:47 INFO memory.MemoryStore: MemoryStore started with capacity 93.3 MB
18/07/11 12:42:47 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/07/11 12:42:48 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers
18/07/11 12:42:48 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (4096 MB per container)
18/07/11 12:42:48 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
18/07/11 12:42:48 INFO yarn.Client: Setting up container launch context for our AM
18/07/11 12:42:48 INFO yarn.Client: Setting up the launch environment for our AM container
18/07/11 12:42:48 INFO yarn.Client: Preparing resources for our AM container
18/07/11 12:42:49 INFO yarn.Client: Uploading resource file:/tmp/spark-f93c6396-2721-4872-9f75-1d2990e95295/__spark_conf__2906525085[illegible]1817990.zip -> hdfs://nameser
vice1/user/edureka_321047/.sparkStaging/application_1528714825862_11994/__spark_conf__.zip
18/07/11 12:42:49 INFO spark.SecurityManager: Changing view acls to: edureka_321047
18/07/11 12:42:49 INFO spark.SecurityManager: Changing modify acls to: edureka_321047
18/07/11 12:42:49 INFO spark.SecurityManager: Changing view acls groups to:
18/07/11 12:42:49 INFO spark.SecurityManager: Changing modify acls groups to:
18/07/11 12:42:49 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acl
 with view permissions: Set(); users  with modify permissions: Set(edureka_321047); groups wit
18/07/11 12:42:49 INFO yarn.Client: Submitting application application_1528714825862_11994 to
18/07/11 12:42:49 INFO impl.YarnClientImpl: Submitted application application_1528714825862_11
18/07/11 12:42:49 INFO cluster.SchedulerExtensionServices: Starting Yarn extension services wi
18/07/11 12:42:50 INFO yarn.Client: Application report for application_1528714825862_11994 (st
18/07/11 12:42:50 INFO yarn.Client:
```

Once you run the ***spark2-submit*** command the entire operation will be pulled and run over spark shell. You can now go back to Hue and verify the output of the operation you have performed.

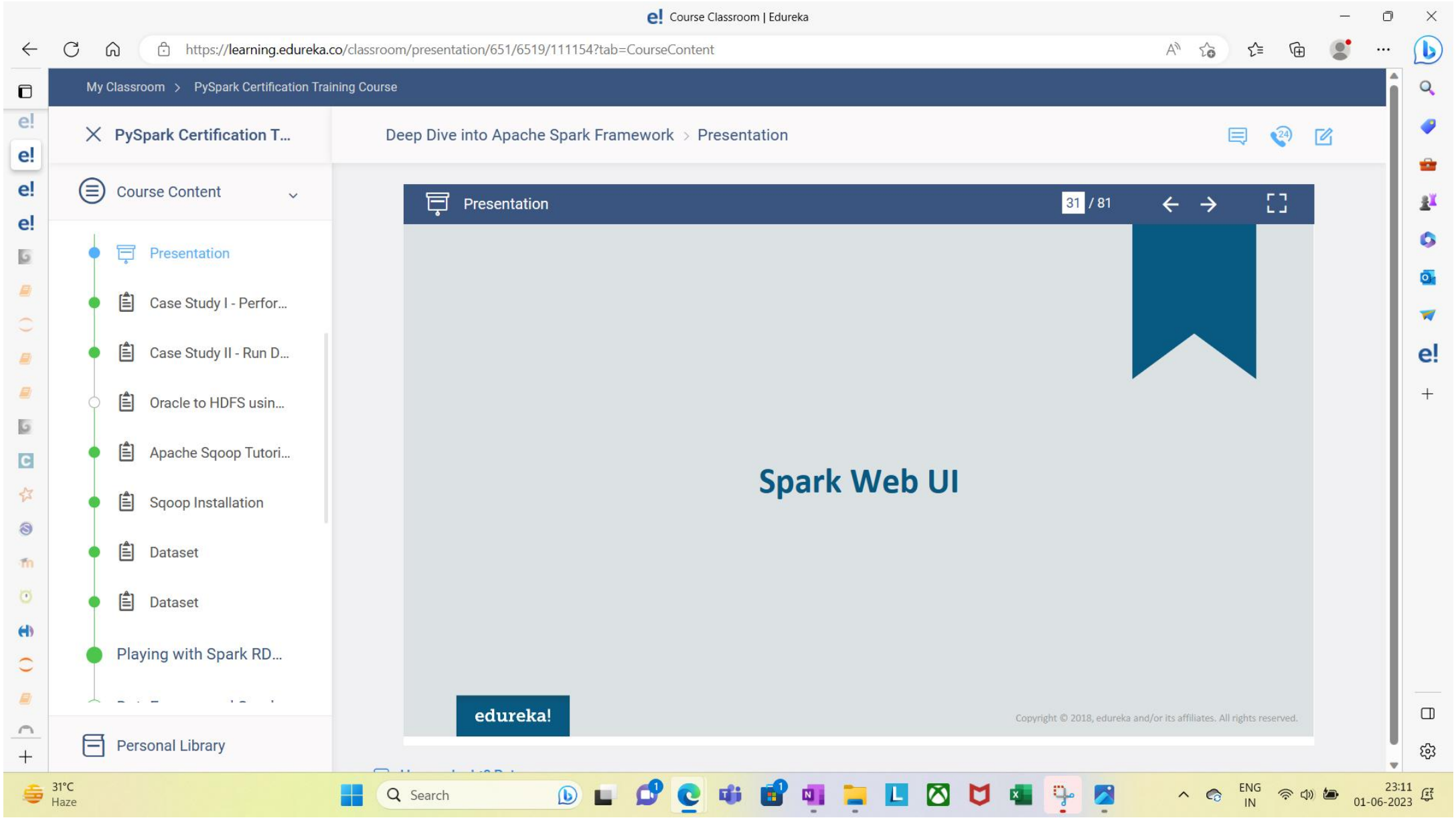
edureka!

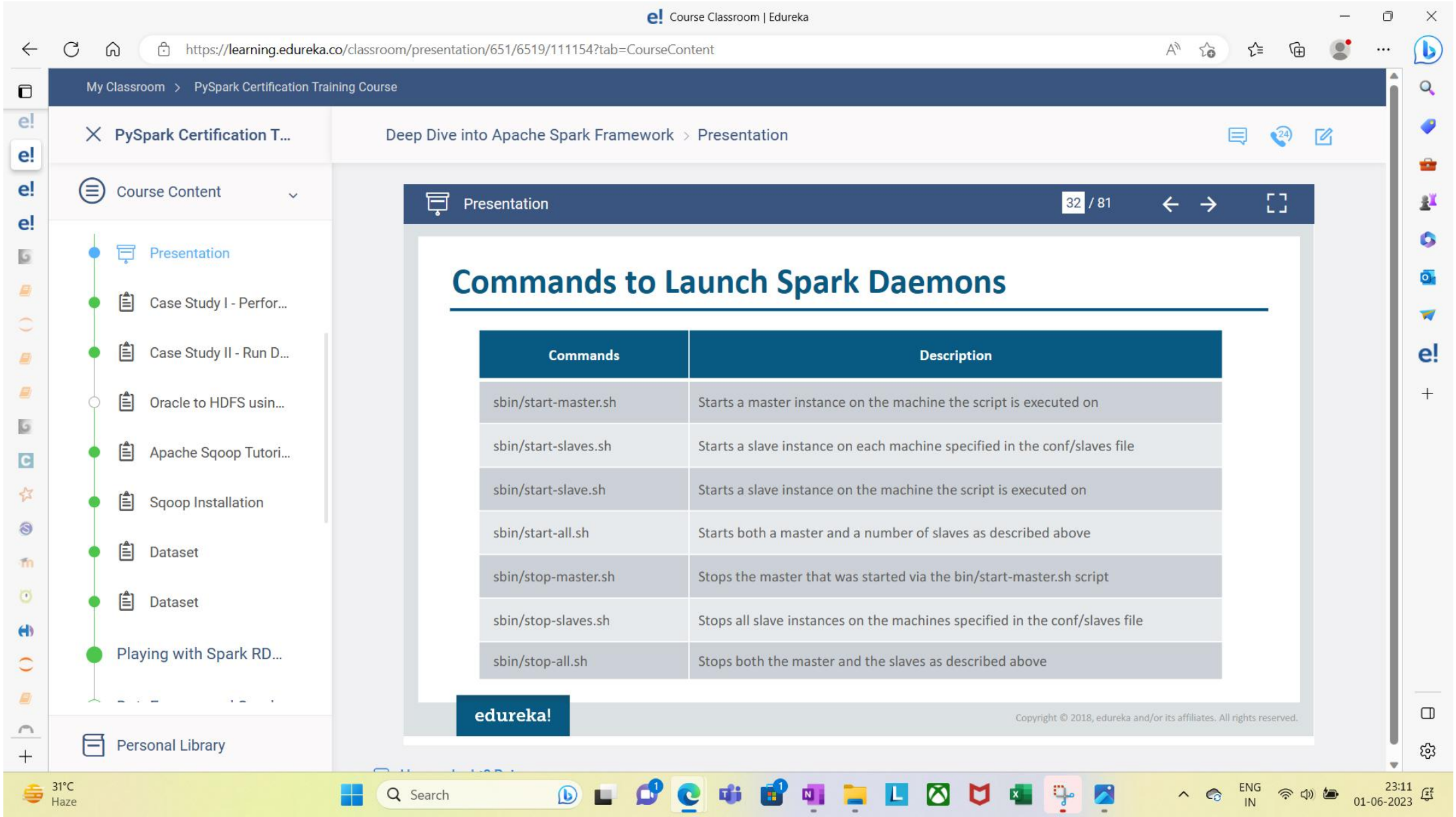
Presentation 30 / 81

# Output of spark-submit Job




edureka!

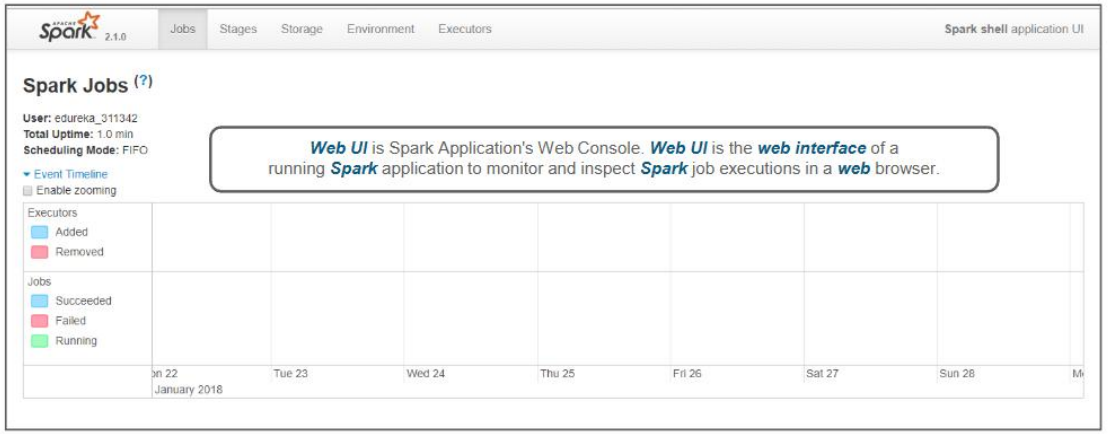

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
31 / 81
Spark Web UI
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
32 / 81
Commands to Launch Spark Daemons
Commands
Description
sbin/start-master.sh
Starts a master instance on the machine the script is executed on
sbin/start-slaves.sh
Starts a slave instance on each machine specified in the conf/slaves file
sbin/start-slave.sh
Starts a slave instance on the machine the script is executed on
sbin/start-all.sh
Starts both a master and a number of slaves as described above
sbin/stop-master.sh
Stops the master that was started via the bin/start-master.sh script
sbin/stop-slaves.sh
Stops all slave instances on the machines specified in the conf/slaves file
sbin/stop-all.sh
Stops both the master and the slaves as described above
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

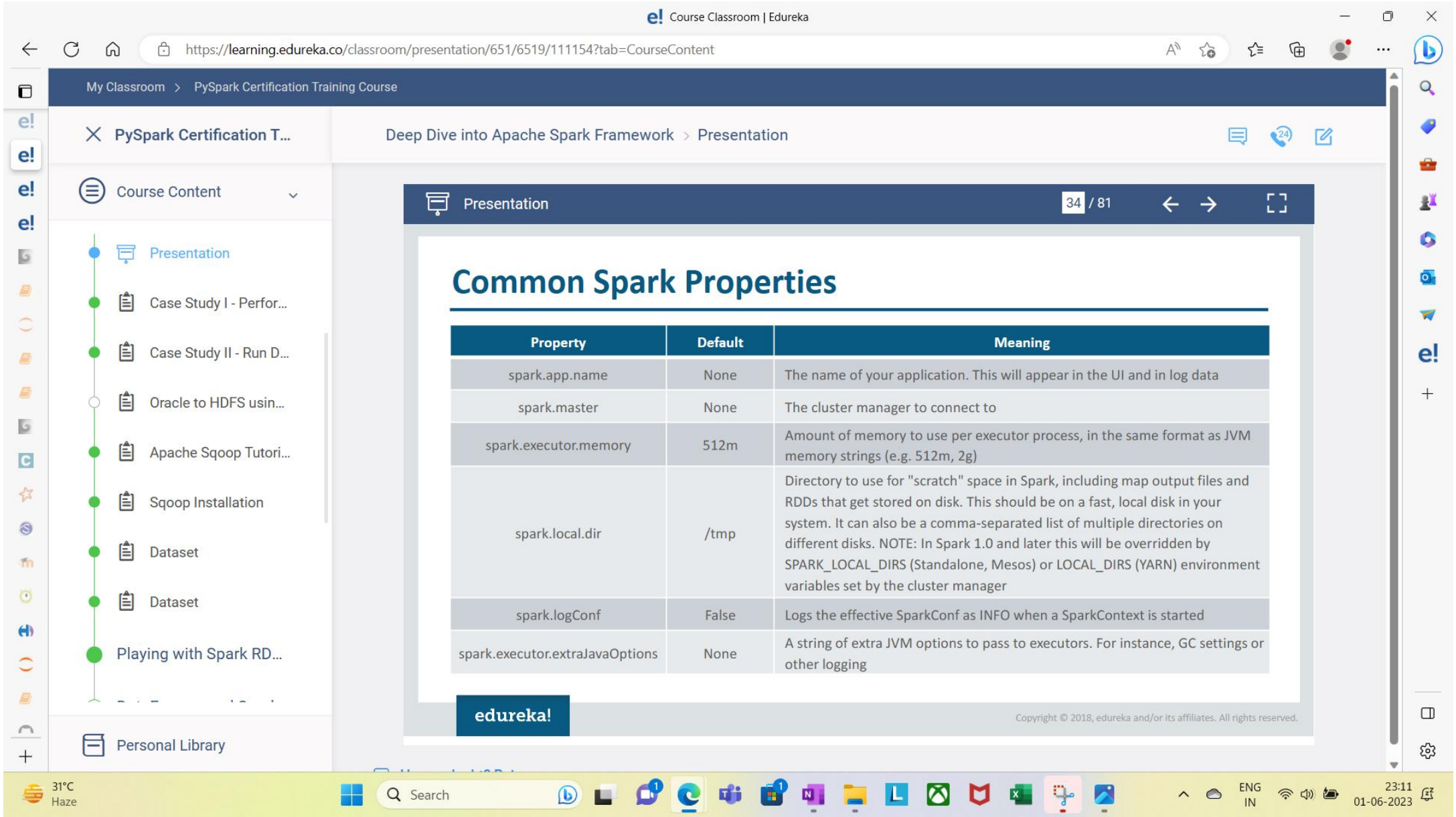
Presentation 33 / 81

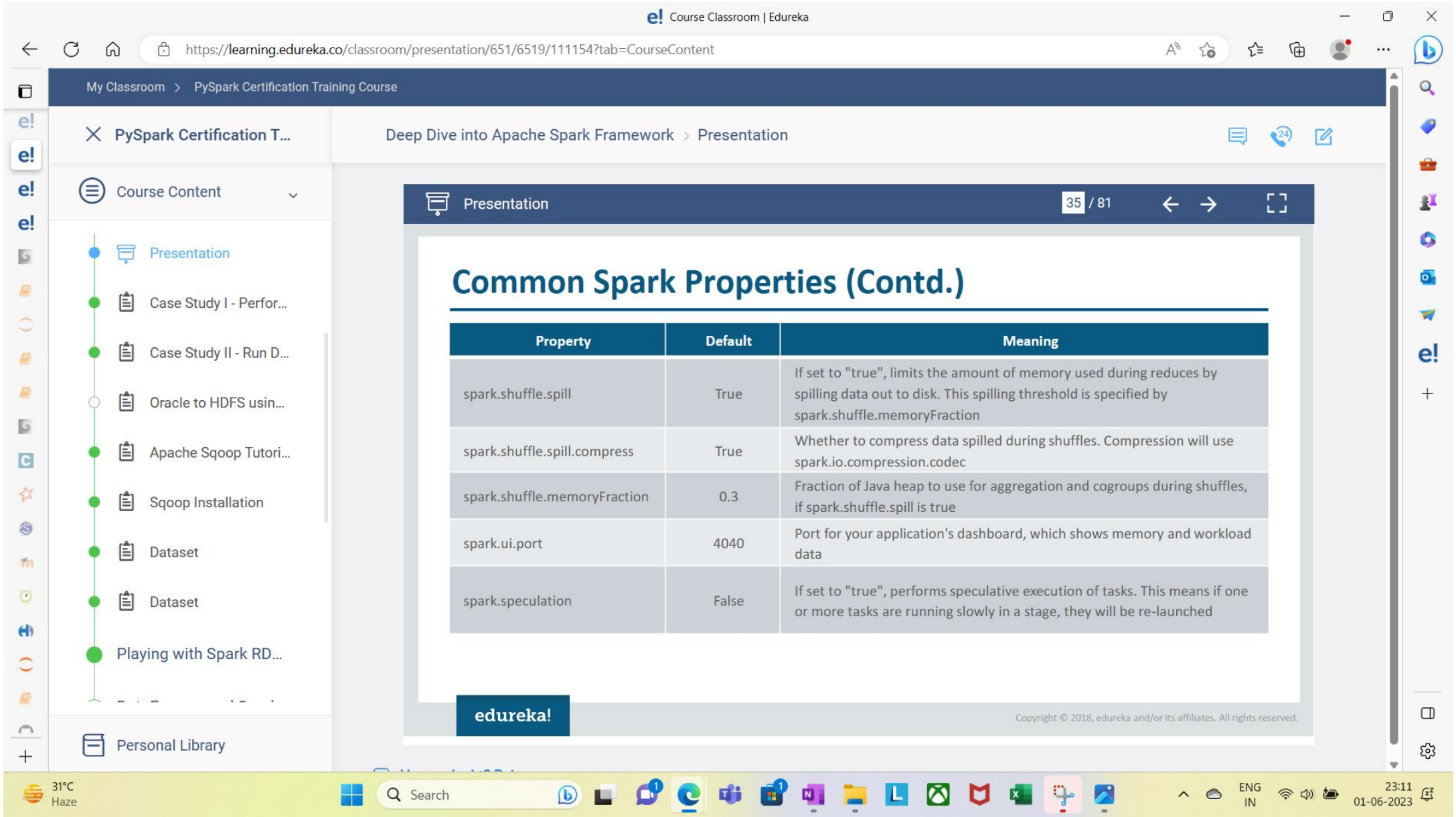
# Spark Web UI

**Go to your Job History Server and Search for your App ID**




edureka!

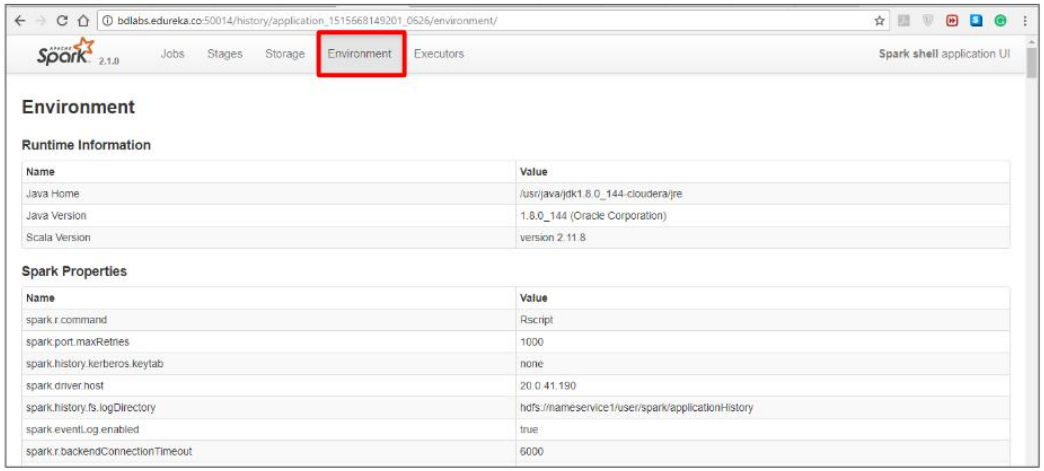

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
34 / 81
Common Spark Properties
Property | Default | Meaning
spark.app.name | None | The name of your application. This will appear in the UI and in log data
spark.master | None | The cluster manager to connect to
spark.executor.memory | 512m | Amount of memory to use per executor process, in the same format as JVM memory strings (e.g. 512m, 2g)
spark.local.dir | /tmp | Directory to use for "scratch" space in Spark, including map output files and RDDs that get stored on disk. This should be on a fast, local disk in your system. It can also be a comma-separated list of multiple directories on different disks. NOTE: In Spark 1.0 and later this will be overridden by SPARK_LOCAL_DIRS (Standalone, Mesos) or LOCAL_DIRS (YARN) environment variables set by the cluster manager
spark.logConf | False | Logs the effective SparkConf as INFO when a SparkContext is started
spark.executor.extraJavaOptions | None | A string of extra JVM options to pass to executors. For instance, GC settings or other logging
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:11 01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
35 / 81
Common Spark Properties (Contd.)
Property
Default
Meaning
spark.shuffle.spill
True
If set to "true", limits the amount of memory used during reduces by spilling data out to disk. This spilling threshold is specified by spark.shuffle.memoryFraction
spark.shuffle.spill.compress
True
Whether to compress data spilled during shuffles. Compression will use spark.io.compression.codec
spark.shuffle.memoryFraction
0.3
Fraction of Java heap to use for aggregation and cogroups during shuffles, if spark.shuffle.spill is true
spark.ui.port
4040
Port for your application's dashboard, which shows memory and workload data
spark.speculation
False
If set to "true", performs speculative execution of tasks. This means if one or more tasks are running slowly in a stage, they will be re-launched
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

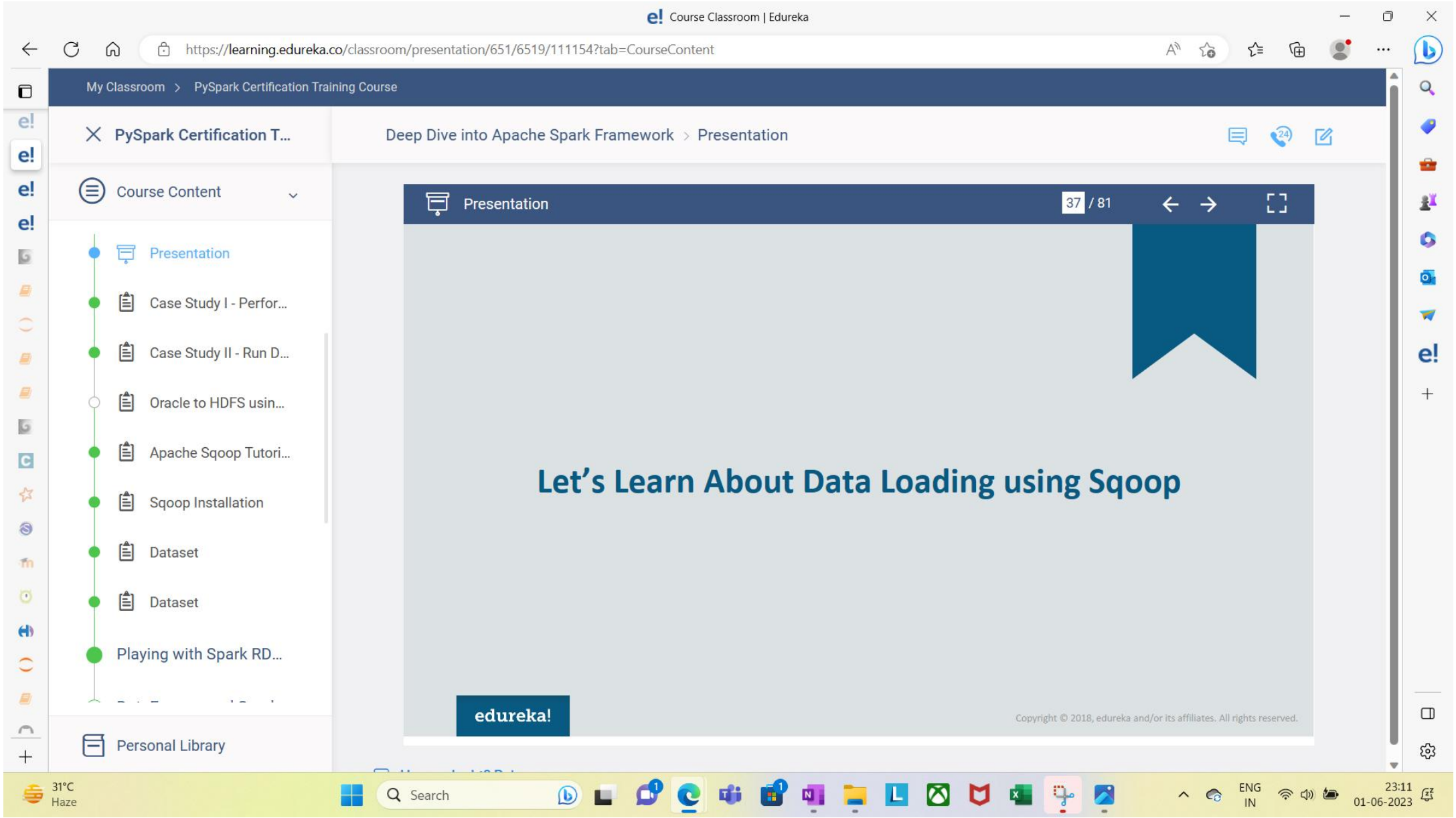
Presentation 36 / 81

# Viewing Spark Properties (Job History Server)

The Environment tab shows the current Spark configuration settings.




edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
37 / 81
Let's Learn About Data Loading using Sqoop
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

Presentation 38 / 81

# Why Sqoop?

**Problems with relational databases:**

- *Structured data stored in databases* and EDW is not easily accessible for analysis in Hadoop
- RDBMS finds it *challenging to handle huge data* volumes
- Relational databases *can't store unstructured data*
- Transferring data using scripts is an inefficient and time consuming task

edureka!

Presentation 39 / 81

# Why Sqoop?

**Solution to the problem using Sqoop:**

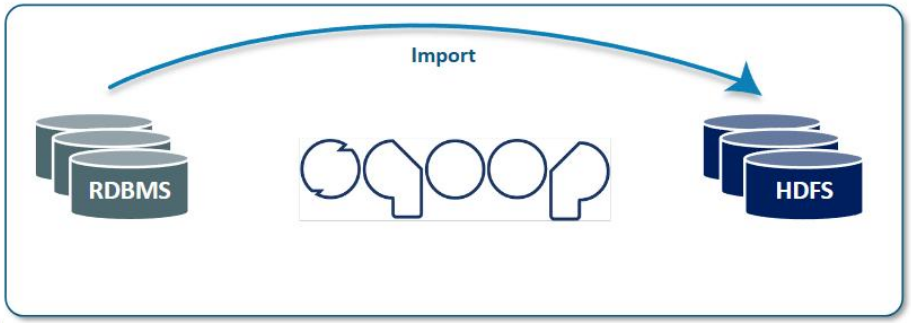
- Sqoop parallelizes *data transfer* for *fast performance* and *optimal system utilization*
- Makes *data analysis more efficient*

edureka!

Presentation 40 / 81

# What is Sqoop?

***Apache Sqoop*** is a tool designed for *efficiently transferring bulk data* between *Apache Hadoop* and *structured data stores* such as relational databases like *Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB* etc.




edureka!

# What is Sqoop?

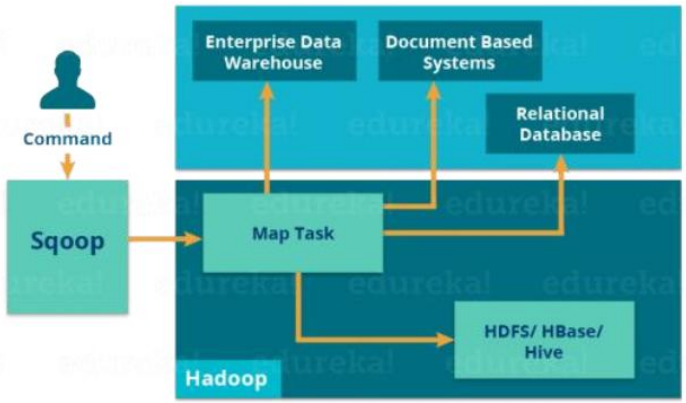
*Sqoop helps offload certain (such as ETL processing) from the EDW to Hadoop for efficient execution at much lower cost. In a broad sense you can import and export data from a regular RDBMS to a HDFS.*


edureka!

Presentation 42 / 81

# Sqoop Architecture

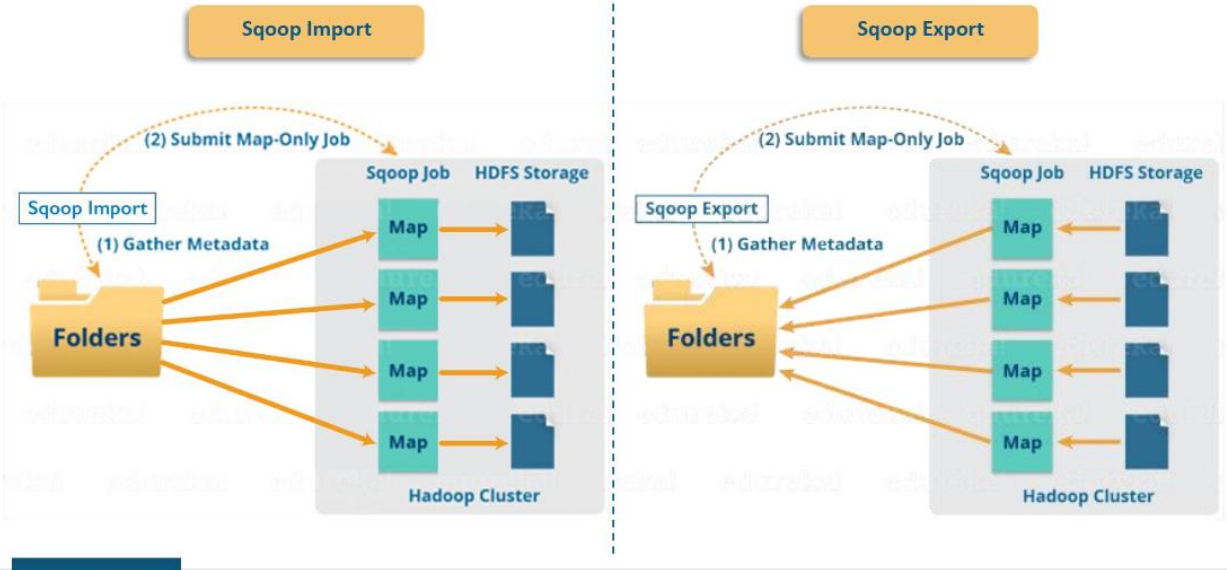
- When we submit Sqoop command, our *main task gets divided into sub tasks* which is *handled by individual Map Task internally*
- Map Task is the sub task, which imports part of data to the Hadoop Ecosystem. Collectively, all Map tasks imports the whole data




edureka!

Presentation 43 / 81

## How Sqoop Import & Export Works?




edureka!

Presentation 44 / 81

# Sqoop Commands

```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop help
Warning: /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/26 06:03:05 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.11.1
usage: sqoop COMMAND [ARGS]

Available commands:
  codegen            Generate code to interact with database records
  create-hive-table  Import a table definition into Hive
  eval               Evaluate a SQL statement and display the results
  export             Export an HDFS directory to a database table
  help               List available commands
  import             Import a table from a database to HDFS
  import-all-tables  Import tables from a database to HDFS
  import-mainframe   Import datasets from a mainframe server to HDFS
  job                Work with saved jobs
  list-databases     List available databases on a server
  list-tables        List available tables in a database
  merge              Merge results of incremental imports
  metastore          Run a standalone Sqoop metastore
  version            Display version information

See 'sqoop help COMMAND' for information on a specific command.
[edureka_321047@ip-20-0-41-202 ~]$
```

edureka!

Presentation 45 / 81

# Login into MySQL

```
[edureka_321047@ip-20-0-41-202 ~]$ mysql -h mysqldb.edu.cloudlab.com -u labuser -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 22316
Server version: 5.7.21 MySQL Community Server (GPL)

Copyright (c) 2000, 2017, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]>
```

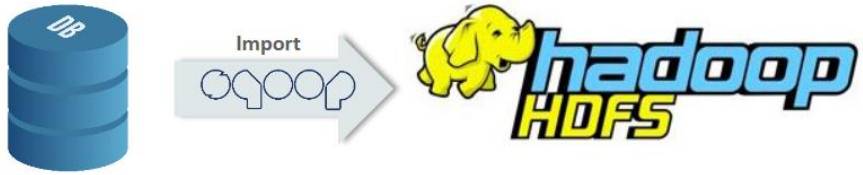
Used to login into MySQL through Webconsole

edureka!

Presentation 46 / 81

# Sqoop Import (MySQL to HDFS)

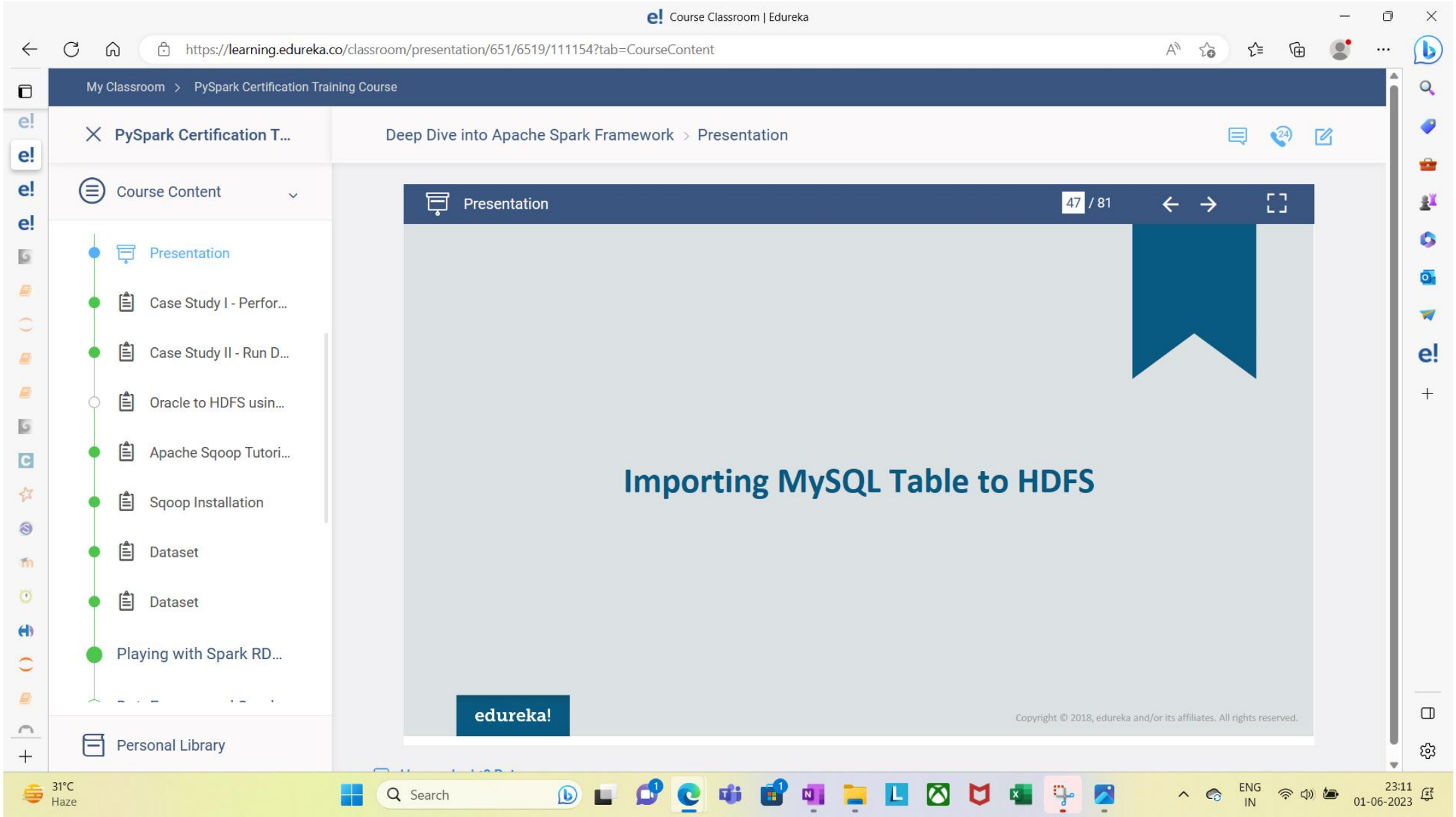
- The *Sqoop import tool* will import each table of the RDBMS in Hadoop.




Command:

```
$ sqoop import \
  --connect jdbc:mysql://<ip address>/<database name>
  --table <mysql_table name>
  --username <username_for_mysql_user> --password <Password>
  -m <number of mappers to run>
  --target-dir <target directory where data needs to be imported>
```

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
47 / 81
Importing MySQL Table to HDFS
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

# Importing MySQL Table to HDFS

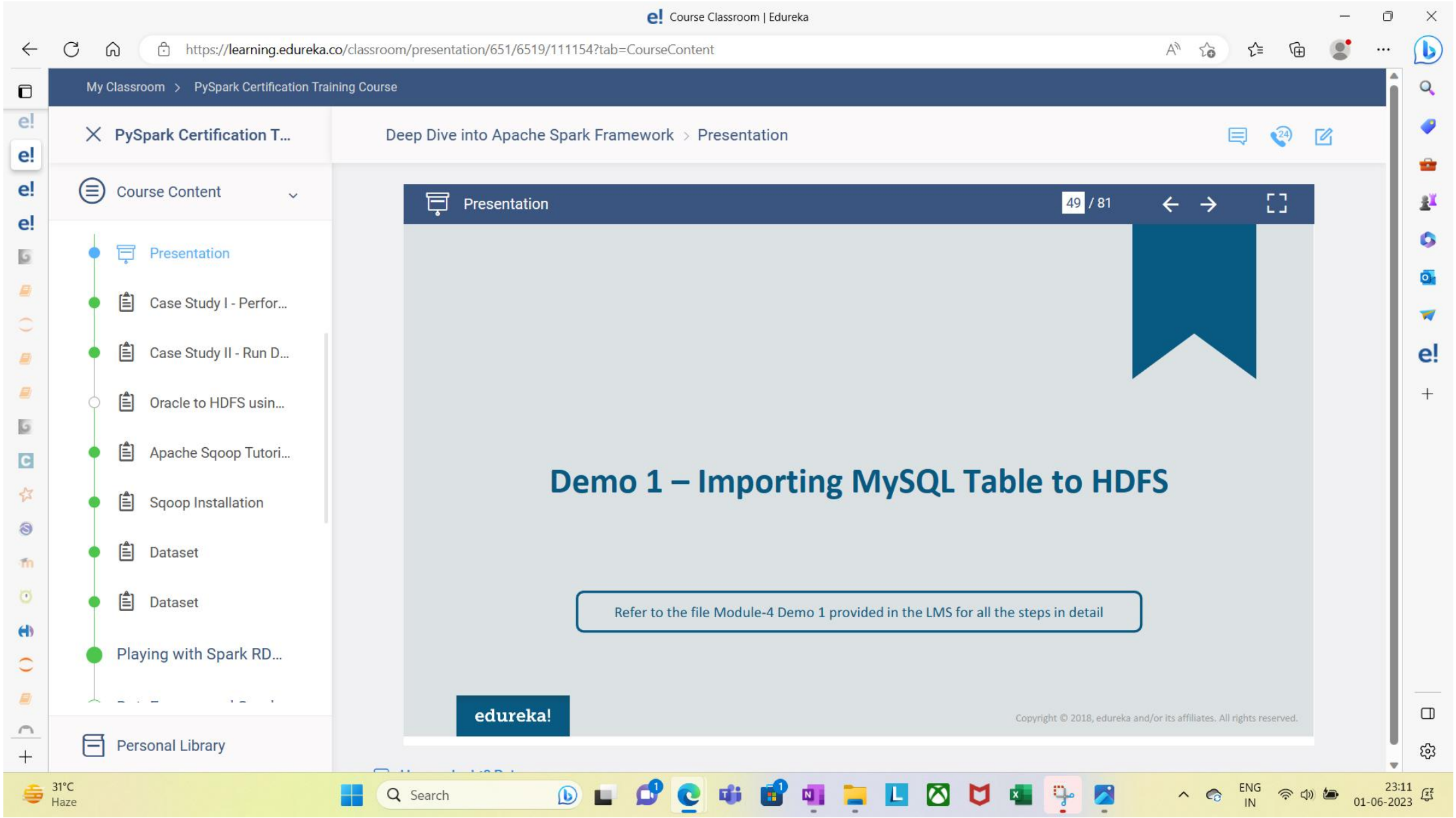
- Let us look at *Sqoop Import function* using a simple example here
- For example, we have to import the below table to HDFS
- The table name is "*CUSTOMERS*" in the database "*Sqoop_Demo*" in MySQL database server

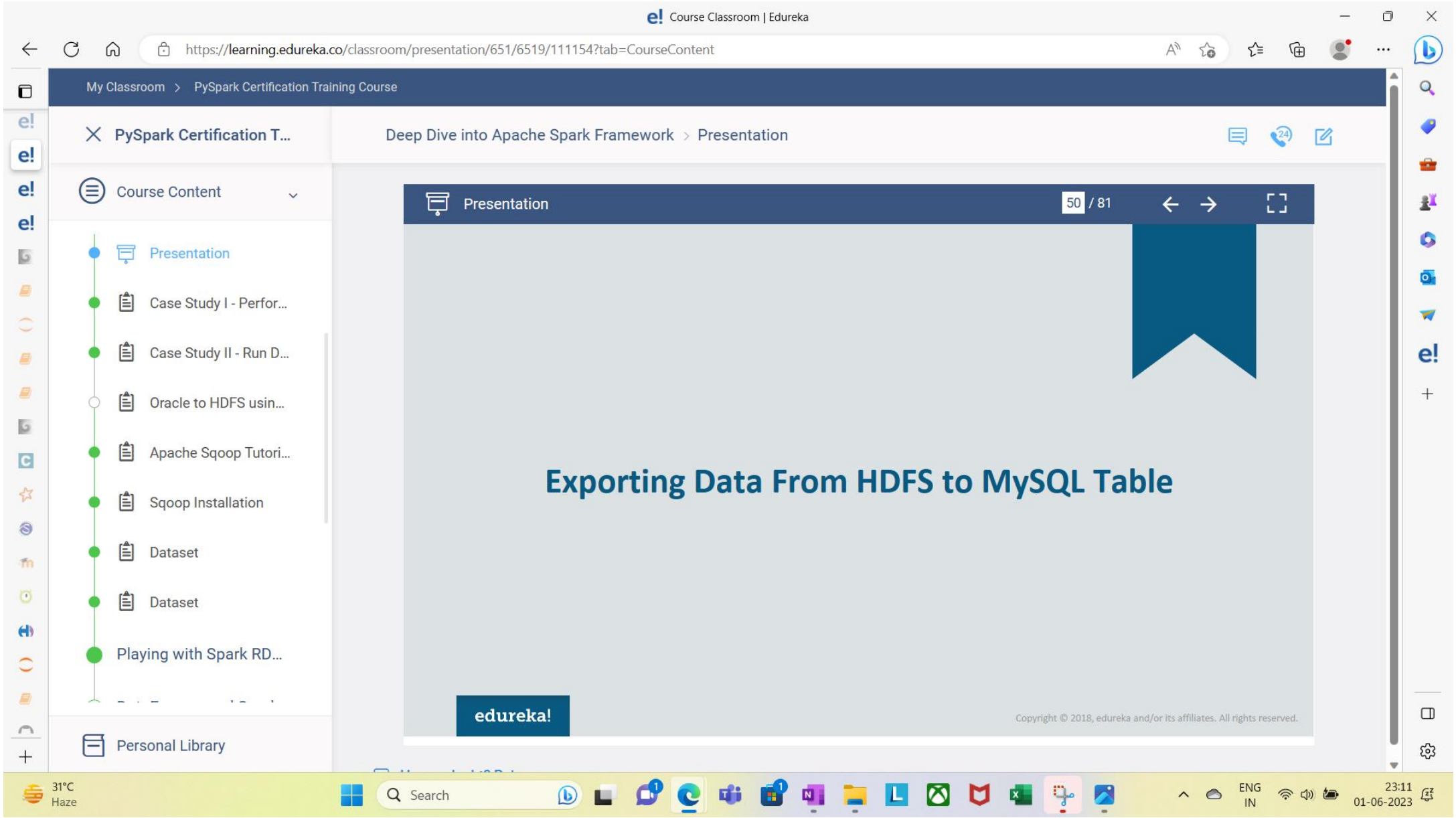
| ID | NAME | AGE | ADDRESS | SALARY |
|---|---|---|---|---|
| 1 | Ramesh | 32 | Ahmedabad | 2000.00 |
| 2 | khilan | 25 | Delhi | 1500.00 |
| 3 | Kaushik | 23 | Kota | 2000.00 |
| 4 | Chaitali | 25 | Mumbai | 6500.00 |
| 5 | Hardik | 27 | Bhopal | 8500.00 |
| 6 | Komal | 22 | MP | 4500.00 |
| 7 | Muffy | 24 | Indore | 10000.00 |

Command to import the *CUSTOMERS* table from MySQL database server to HDFS.

```
$ sqoop import --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo --
username labuser --password edureka --table CUSTOMERS -m 1 --target-dir
'/user/edureka_249489/sqoop_sql/';
```

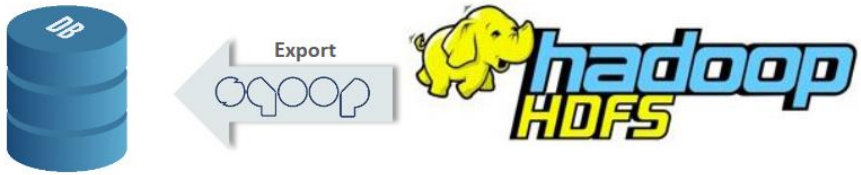
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
49 / 81
Demo 1 – Importing MySQL Table to HDFS
Refer to the file Module-4 Demo 1 provided in the LMS for all the steps in detail
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
50 / 81
Exporting Data From HDFS to MySQL Table
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

# Sqoop Export (HDFS to MySQL)

- Sqoop provides the feature to *export data* from *HDFS to RDBMS*




**Command:**

```
$ sqoop export \
  --connect jdbc:mysql://<ip address>/<database name>
  --table <mysql_table name>
  --username <username_for_mysql_user> --password <Password>
  -m <number of mappers to run>
  --export-dir <directory name from where data needs to be exported>
```

edureka!

# Exporting Data to MySQL

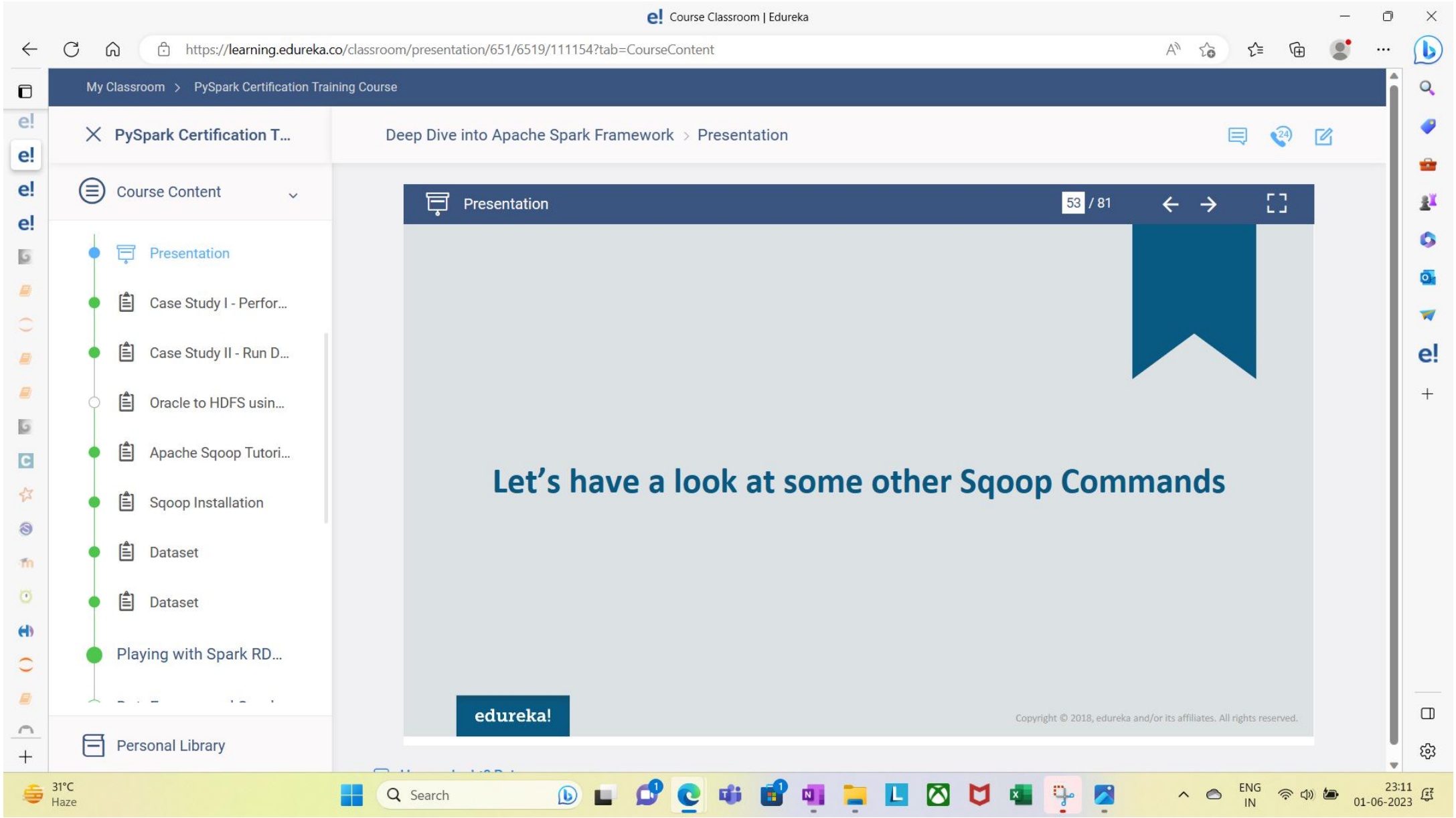
```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop export \
> --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo \
> --table CUSTOMERS \
> --username labuser --password edureka \
> -m 1 \
> --export-dir '/user/edureka_321047/sqoop_sql' \
> ;
```

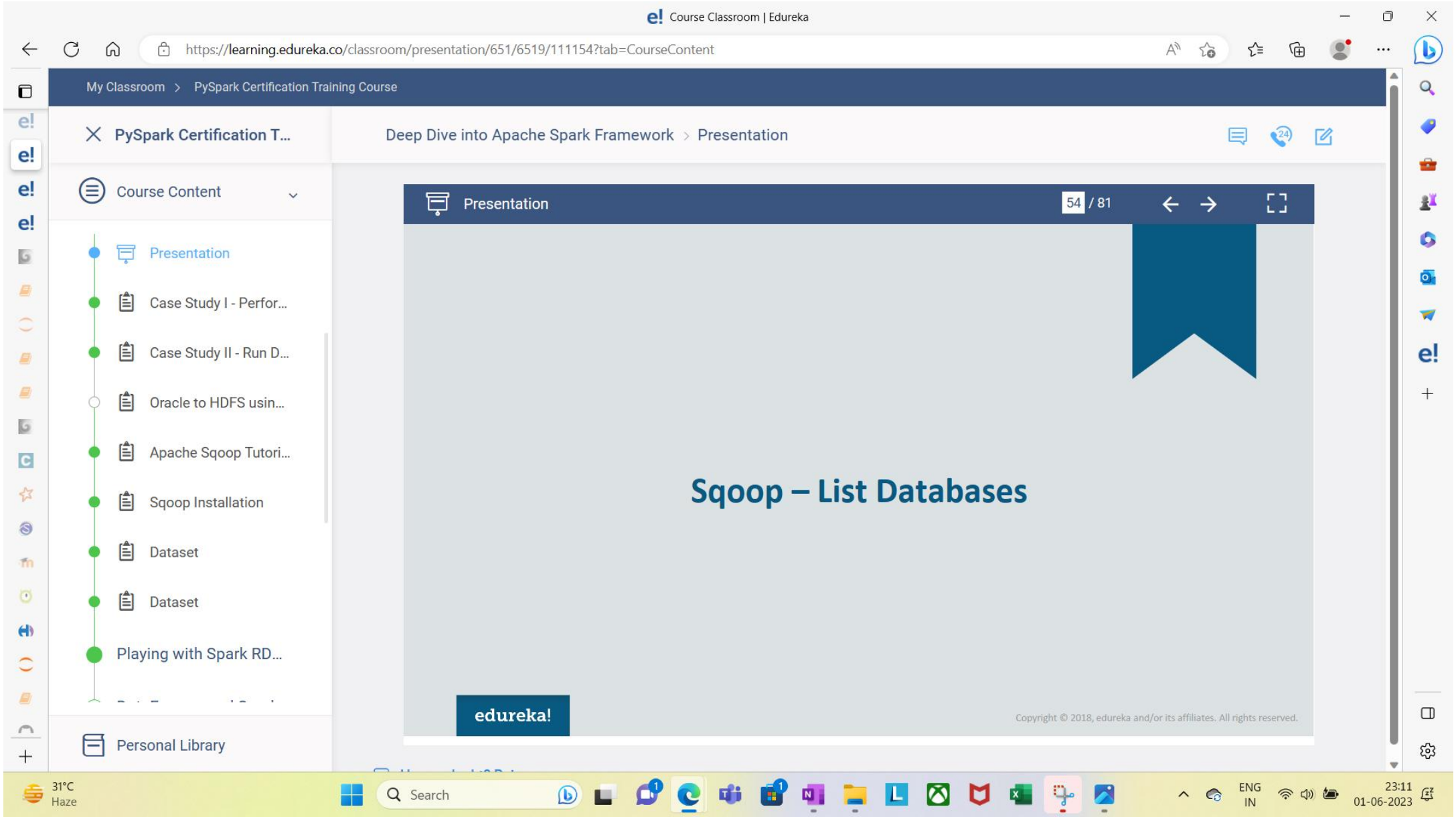
- Considering the same example that we consider for importing the data to HDFS
- It is mandatory that the table to be exported is created manually and is present in the database

```
MySQL [Sqoop_Demo]> delete from CUSTOMERS    // Delete the table 'emp'
    -> ;
Query OK, 7 rows affected (0.00 sec)

MySQL [Sqoop_Demo]> select * from CUSTOMERS
    -> ;
Empty set (0.00 sec)

MySQL [Sqoop_Demo]> select * from CUSTOMERS
    -> ;
+----+----------+-----+-----------+----------+
| ID | NAME     | AGE | ADDRESS   | SALARY   |
+----+----------+-----+-----------+----------+
|  1 | Ramesh   |  32 | Ahmedabad |  2000.00 |
|  2 | Khilan   |  25 | Delhi     |  1500.00 |
|  3 | kaushik  |  23 | Kota      |  2000.00 |
|  4 | Chaitali |  25 | Mumbai    |  6500.00 |
|  5 | Hardik   |  27 | Bhopal    |  8500.00 |
|  6 | Komal    |  22 | MP        |  4500.00 |
|  7 | Muffy    |  24 | Indore    | 10000.00 |
+----+----------+-----+-----------+----------+
7 rows in set (0.00 sec)                    // Check data in the table 'emp'

MySQL [Sqoop_Demo]>
```

*Output after running the export command*

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
53 / 81
Let's have a look at some other Sqoop Commands
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
54 / 81
Sqoop – List Databases
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

Presentation 55 / 81

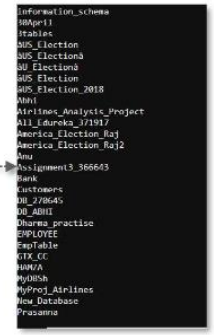
# Sqoop - List Databases

We can list out the databases present in relation database using Sqoop

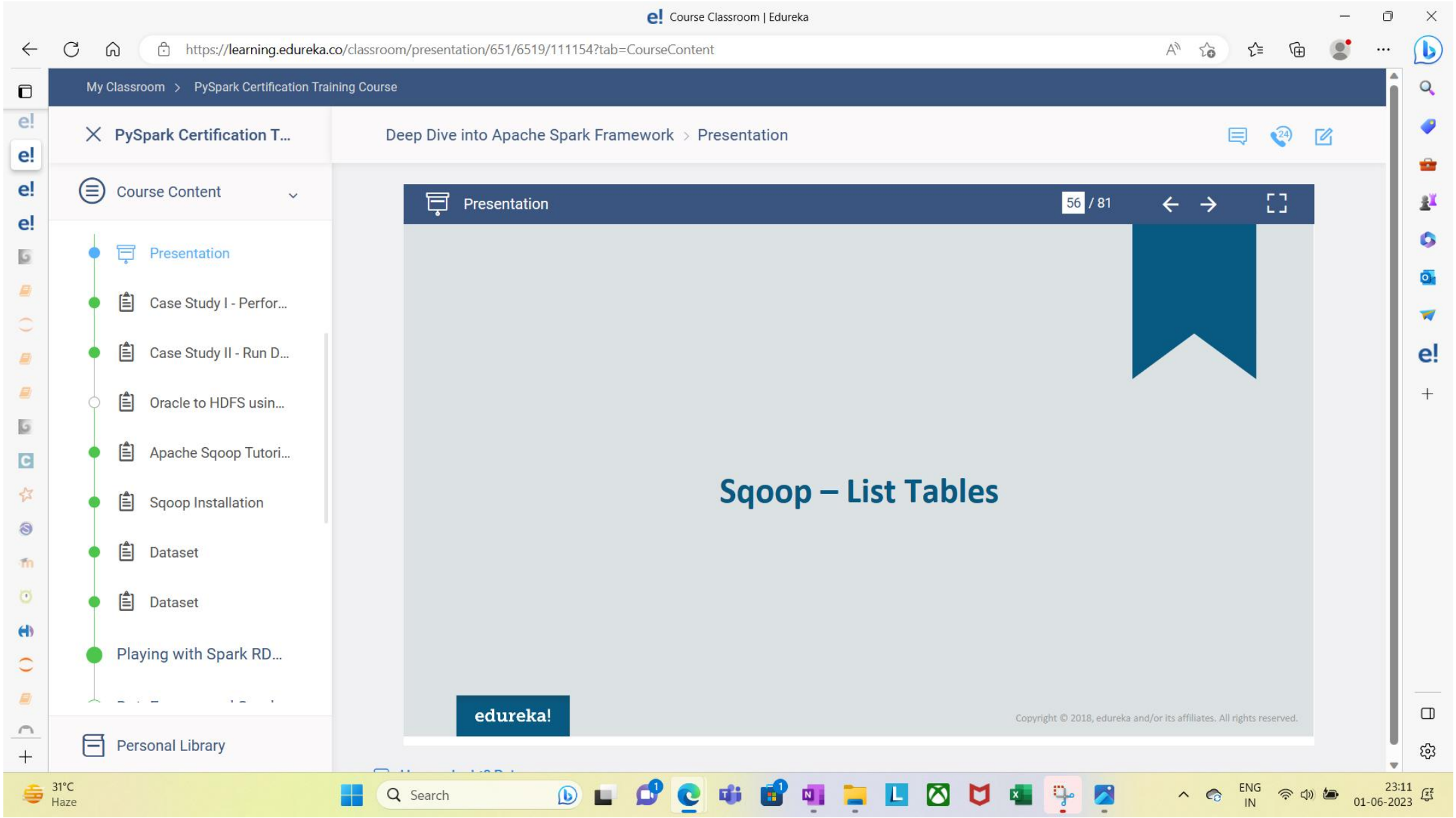
**Command:** `sqoop list-databases --connect jdbc:mysql://sqoopdb.edu.cloudlab.com --username labuser --password edureka`

```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop list-databases --connect jdbc:mysql://sqoopdb.edu.cloudlab.com --username labuser --password edureka
Warning: /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
```

List of Databases




edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
56 / 81
Sqoop – List Tables
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

Presentation 57 / 81

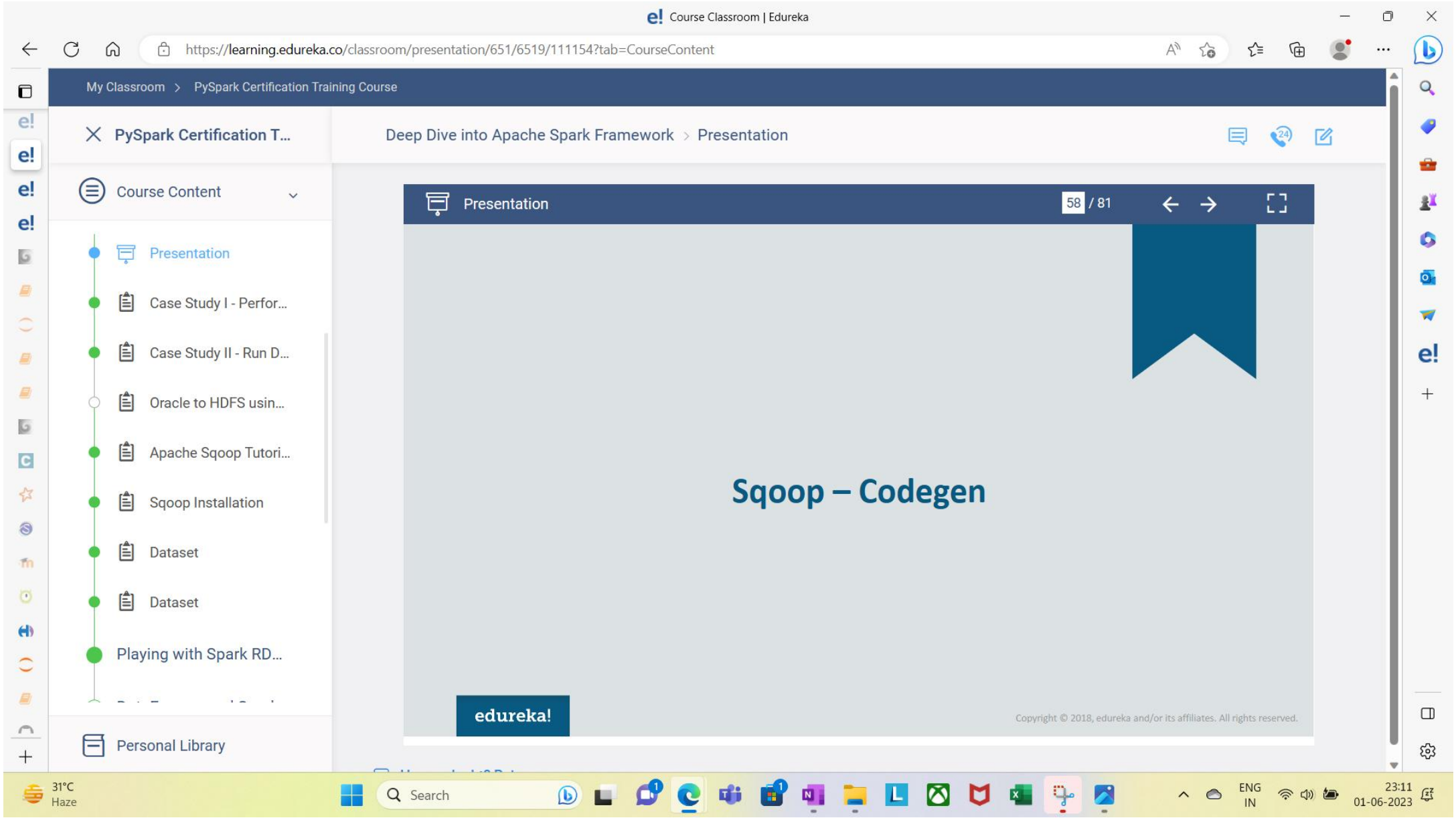
# Sqoop - List Tables

We can also list out the tables of a particular database in MySQL database server using Sqoop

**Command:**
```
sqoop list-tables --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo
--username labuser --password edureka
```

```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop list-tables --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo --username labuser --password edureka
Warning: /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/26 06:57:00 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.11.1
18/06/26 06:57:00 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/06/26 06:57:00 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Tue Jun 26 06:57:00 UTC 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+,
.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SS
rverCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststo
certificate verification.
CUSTOMERS
```

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
58 / 81
Sqoop – Codegen
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:11
01-06-2023

Presentation 59 / 81

# Sqoop - Codegen

In object-oriented application, every database table has one Data Access Object class that contains 'getter' and 'setter' methods to initialize objects.

edureka!

Presentation 60 / 81

# Sqoop - Codegen

In object-oriented application, every database table has one Data Access Object class that contains 'getter' and 'setter' methods to initialize objects.

Codegen generates the DAO class automatically.

edureka!

Presentation 61 / 81

# Sqoop - Codegen

In object-oriented application, every database table has one Data Access Object class that contains 'getter' and 'setter' methods to initialize objects.

Sqoop-Codegen command generates Java class files which encapsulate and interpret imported records.

Codegen generates the DAO class automatically.

edureka!

# Sqoop - Codegen

In object-oriented application, every database table has one Data Access Object class that contains 'getter' and 'setter' methods to initialize objects.

Sqoop-Codegen command generates Java class files which encapsulate and interpret imported records.

The Java definition of a record is initiated as part of the import process

Codegen generates the DAO class automatically.

edureka!

Presentation 63 / 81

# Sqoop - Codegen

In object-oriented application, every database table has one Data Access Object class that contains 'getter' and 'setter' methods to initialize objects.

Sqoop-Codegen command generates Java class files which encapsulate and interpret imported records.

The Java definition of a record is initiated as part of the import process

For example, if Java source is lost, it can be recreated. New versions of a class can be created which use different delimiters between fields, and so on
It generates DAO class in Java, based on the Table Schema structure.

Codegen generates the DAO class automatically.

edureka!

Presentation 64 / 81

# Sqoop - Codegen

The command for generating java code is:

**Command:** `sqoop codegen --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo --username labuser --password edureka --table CUSTOMERS`

```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop codegen --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo --username labuser --password edureka --table CUSTOMERS
Warning: /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/26 07:01:15 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.11.1
18/06/26 07:01:15 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/06/26 07:01:15 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/06/26 07:01:15 INFO tool.CodeGenTool: Beginning code generation
Tue Jun 26 07:01:16 UTC 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5
.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifySe
rverCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server
 certificate verification.
18/06/26 07:01:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `CUSTOMERS` AS t LIMIT 1
18/06/26 07:01:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `CUSTOMERS` AS t LIMIT 1
18/06/26 07:01:17 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/soop-edureka_321047/compile/0e7a42bc68245f545647c37616d1de57/CUSTOMERS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/06/26 07:01:19 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-edureka_321047/compile/0e7a42bc68245f545647c37616d1de57/CUSTOMERS.jar
```

You can see the path in above image where the code is generated. Let us go the path and check the files that are created.

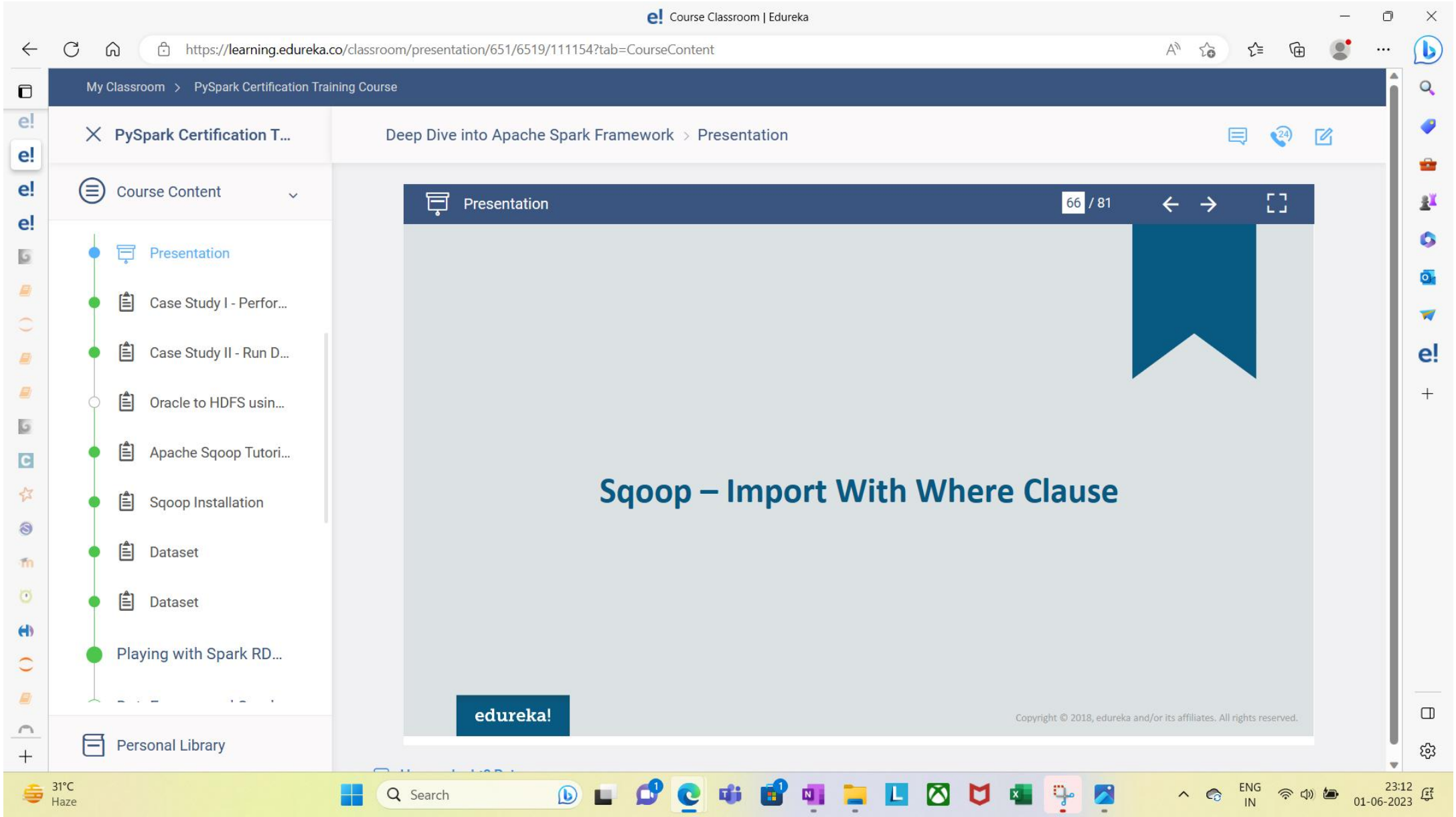
edureka!

Presentation 65 / 81

# Sqoop - Codegen : Generated Code

```
[edureka_249489@ip-20-0-41-93 ~]$ ls /tmp/sqoop-edureka_249489/compile/0fc383f58586f913a85dc8c703df84f1
emp$1.class  emp$2.class  emp$3.class  emp$4.class  emp$5.class  emp.class  emp$FieldSetterCommand.class  emp.jar  emp.java
[edureka_249489@ip-20-0-41-93 ~]$ cat emp.java
// ORM class for table 'emp'
// WARNING: This class is AUTO-GENERATED. Modify at your own risk.
//
// Debug information:
// Generated date: Thu Jan 04 11:35:35 UTC 2018
// For connector: org.apache.sqoop.manager.MySQLManager
import org.apache.hadoop.io.BytesWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.Writable;
import org.apache.hadoop.mapred.lib.db.DBWritable;
import com.cloudera.sqoop.lib.JdbcWritableBridge;
import com.cloudera.sqoop.lib.DelimiterSet;
import com.cloudera.sqoop.lib.FieldFormatter;
import com.cloudera.sqoop.lib.RecordParser;
import com.cloudera.sqoop.lib.BooleanParser;
import com.cloudera.sqoop.lib.BlobRef;
import com.cloudera.sqoop.lib.ClobRef;
import com.cloudera.sqoop.lib.LargeObjectLoader;
import com.cloudera.sqoop.lib.SqoopRecord;
import java.sql.PreparedStatement;
```

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
66 / 81
Sqoop – Import With Where Clause
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:12
01-06-2023

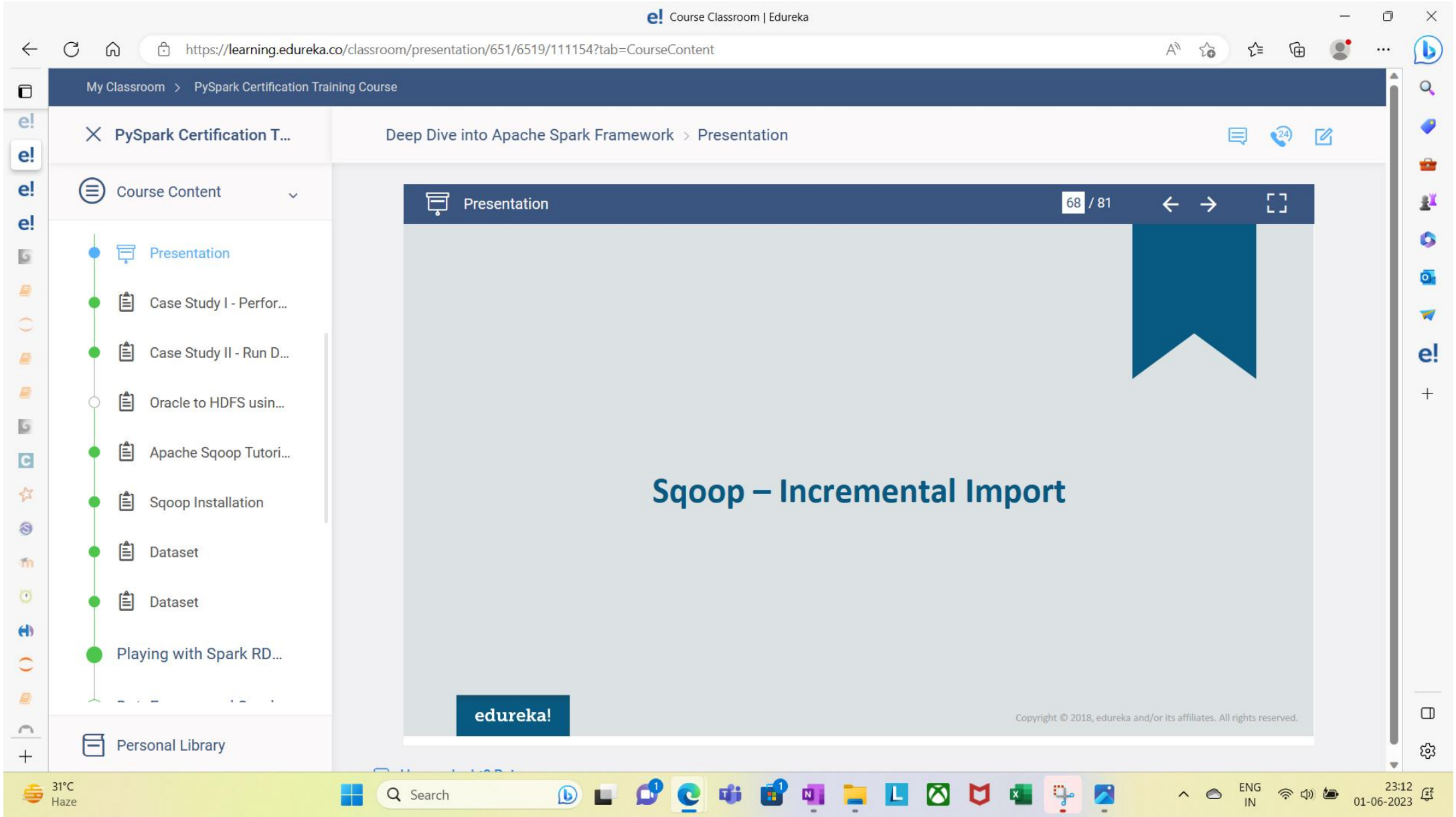
Presentation 67 / 81

# Sqoop - Import With Where Clause

```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop import \
> --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo \
> --table CUSTOMERS \
> --username labuser --password edureka \
> -m 1 \
> --where "Age > 25" \
> --target-dir '/user/edureka_321047/sqoop_sql1/' \
> ;
```

Go to Hue & check the output


edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
68 / 81
Sqoop – Incremental Import
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:12
01-06-2023

Presentation 69 / 81

# Sqoop - Incremental Import

Sqoop provides an incremental import mode which can be used to retrieve only rows newer than some previously-imported set of rows.

Sqoop supports two types of incremental imports:

- Append
- Last Modified

edureka!

# Sqoop - Incremental Import

APPEND

LAST MODIFIED

- You can use the --incremental argument to specify the type of incremental import to perform.
- Specify ***append*** mode when importing a table where new rows are continually being added with increasing row id values.
- You specify the column containing the row's id with ***--check-column***.
- Sqoop imports rows where the check column has a value greater than the one specified with ***--last-value***.

edureka!

Presentation 71 / 81

# Sqoop - Incremental Import


- An alternate table update strategy supported by Sqoop is called ***last modified*** mode.
- You should use this when rows of the source table may be updated, and each such update will set the value of a last-modified column to the current timestamp.

edureka!

# Sqoop - Incremental Import

First we are inserting a new row which will be updated in our HDFS.

```
MySQL [Sqoop_Demo]> INSERT INTO CUSTOMERS (ID,NAME,AGE,ADDRESS,SALARY) VALUES (8, 'Abhay', 28, 'Karnataka', 4300.00 );
Query OK, 1 row affected (0.00 sec)

MySQL [Sqoop_Demo]>
```

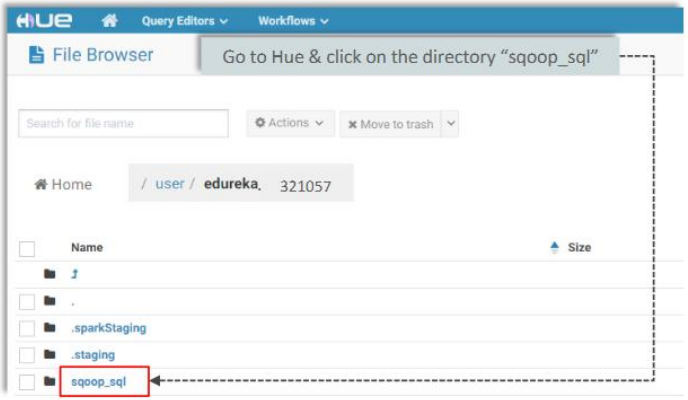
```
MySQL [Sqoop_Demo]> select * from CUSTOMERS;
+----+----------+-----+-----------+----------+
| ID | NAME     | AGE | ADDRESS   | SALARY   |
+----+----------+-----+-----------+----------+
|  1 | Ramesh   |  32 | Ahmedabad |  2000.00 |
|  2 | Khilan   |  25 | Delhi     |  1500.00 |
|  3 | kaushik  |  23 | Kota      |  2000.00 |
|  4 | Chaitali |  25 | Mumbai    |  6500.00 |
|  5 | Hardik   |  27 | Bhopal    |  8500.00 |
|  6 | Komal    |  22 | MP        |  4500.00 |
|  7 | Muffy    |  24 | Indore    | 10000.00 |
|  8 | Abhay    |  28 | Karnataka |  4300.00 |
+----+----------+-----+-----------+----------+
8 rows in set (0.00 sec)
```

edureka!

Presentation 73 / 81

# Sqoop - Incremental Import Command

```
[edureka_321047@ip-20-0-41-202 ~]$ sqoop import --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo --table CUSTOMERS --username labuser --password edureka
--target-dir '/user/edureka_321047/sqoop_sql/' --incremental append --check-column Id --last-value 7
```

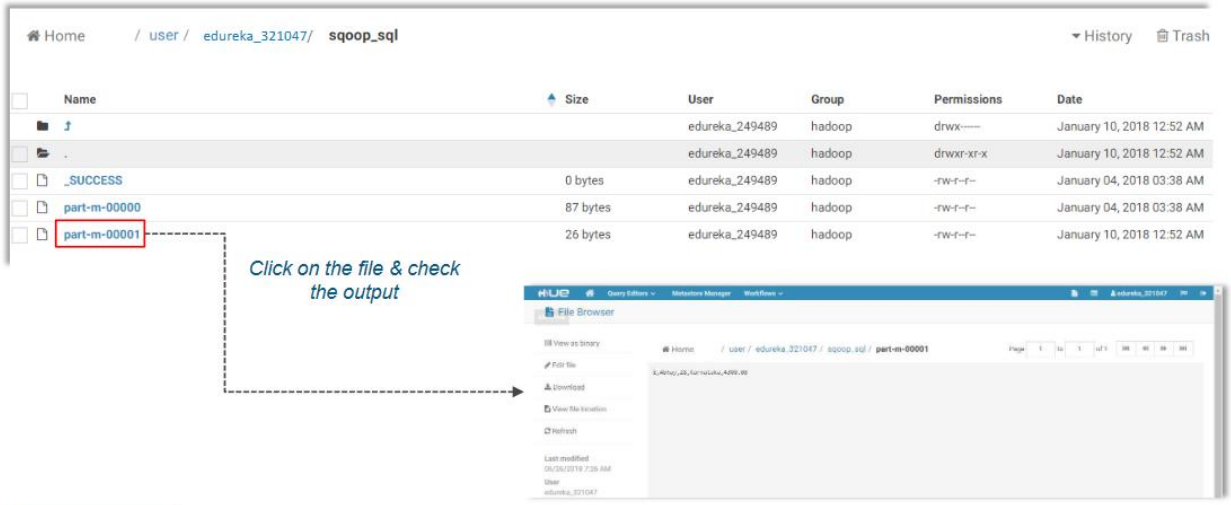



edureka!

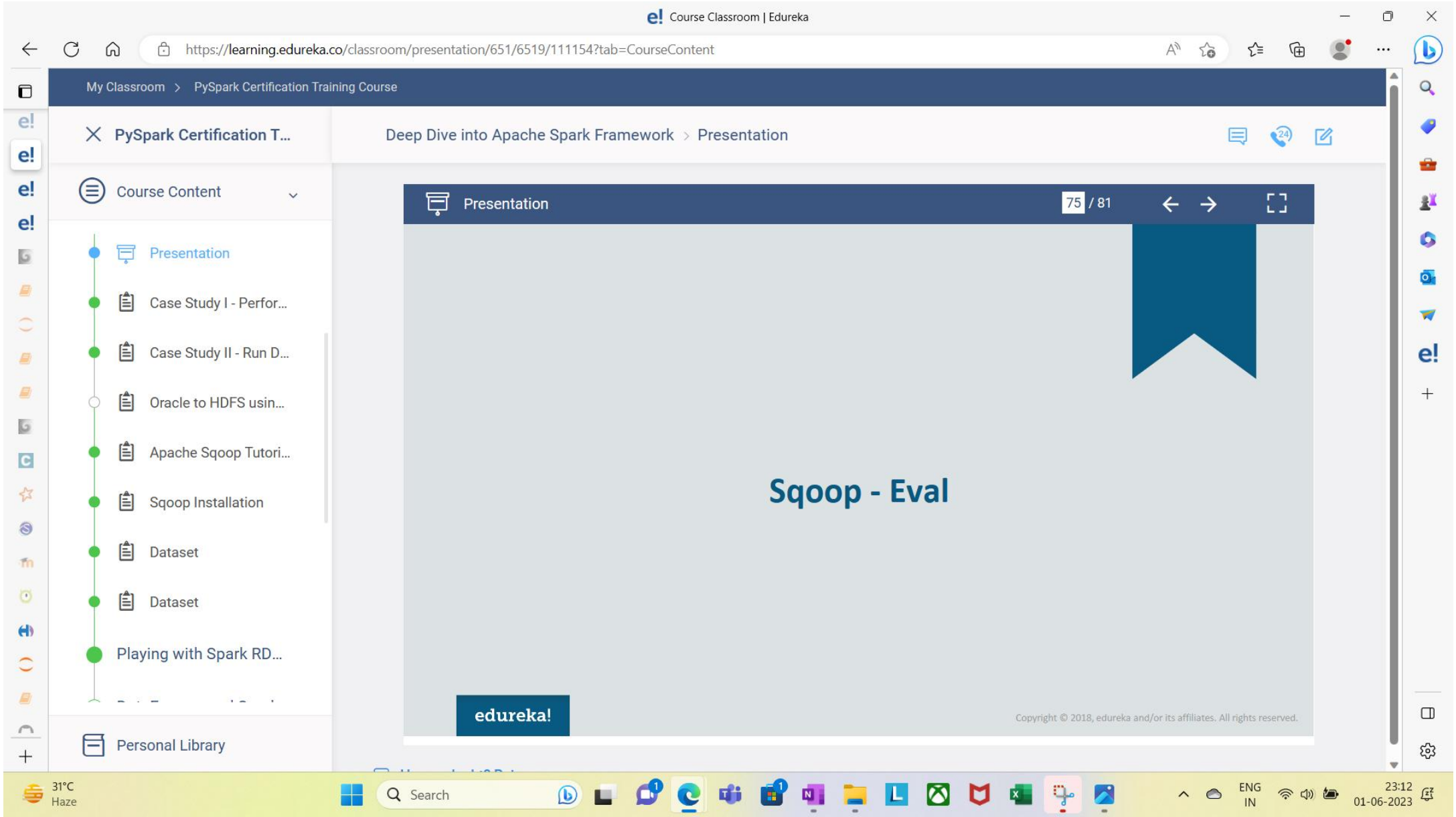
Presentation 74 / 81

# Sqoop - Incremental Import

You can see in the below image, a new file is created with the updated data




edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
75 / 81
Sqoop - Eval
edureka!
Copyright © 2018, edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:12
01-06-2023

# Sqoop - Eval

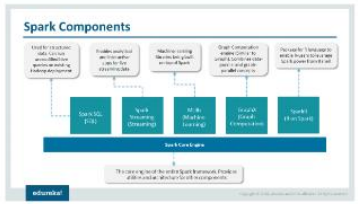
- The eval tool can be used for evaluation purpose only
- The eval tool allows to tun simple SQL queries against the database and the results are printed to console
- You can verify the database connection within Sqoop

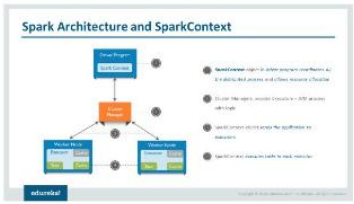
```
[edureka_321047@ip-20-0-32-175 ~]$ sqoop eval --connect jdbc:mysql://sqoopdb.edu.cloudlab.com/Sqoop_Demo --username labuser --password edureka --query "SELECT * FROM CUSTOMERS LIMIT 3"
Warning: /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/26 12:30:59 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.11.1
18/06/26 12:30:59 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/06/26 12:30:59 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Tue Jun 26 12:30:59 UTC 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
-------------------------------------------------------------------------------
| ID          | NAME                | AGE         | ADDRESS             | SALARY              |
-------------------------------------------------------------------------------
| 1           | Ramesh              | 32          | Ahmedabad           | 2000.00             |
| 2           | Khilan              | 25          | Delhi               | 1500.00             |
| 3           | kaushik             | 23          | Kota                | 2000.00             |
-------------------------------------------------------------------------------
[edureka_321047@ip-20-0-32-175 ~]$
```

edureka!

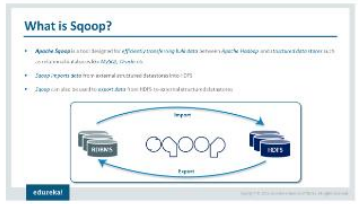
Presentation 77 / 81

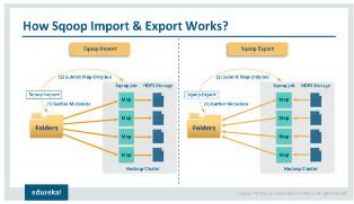
# Summary



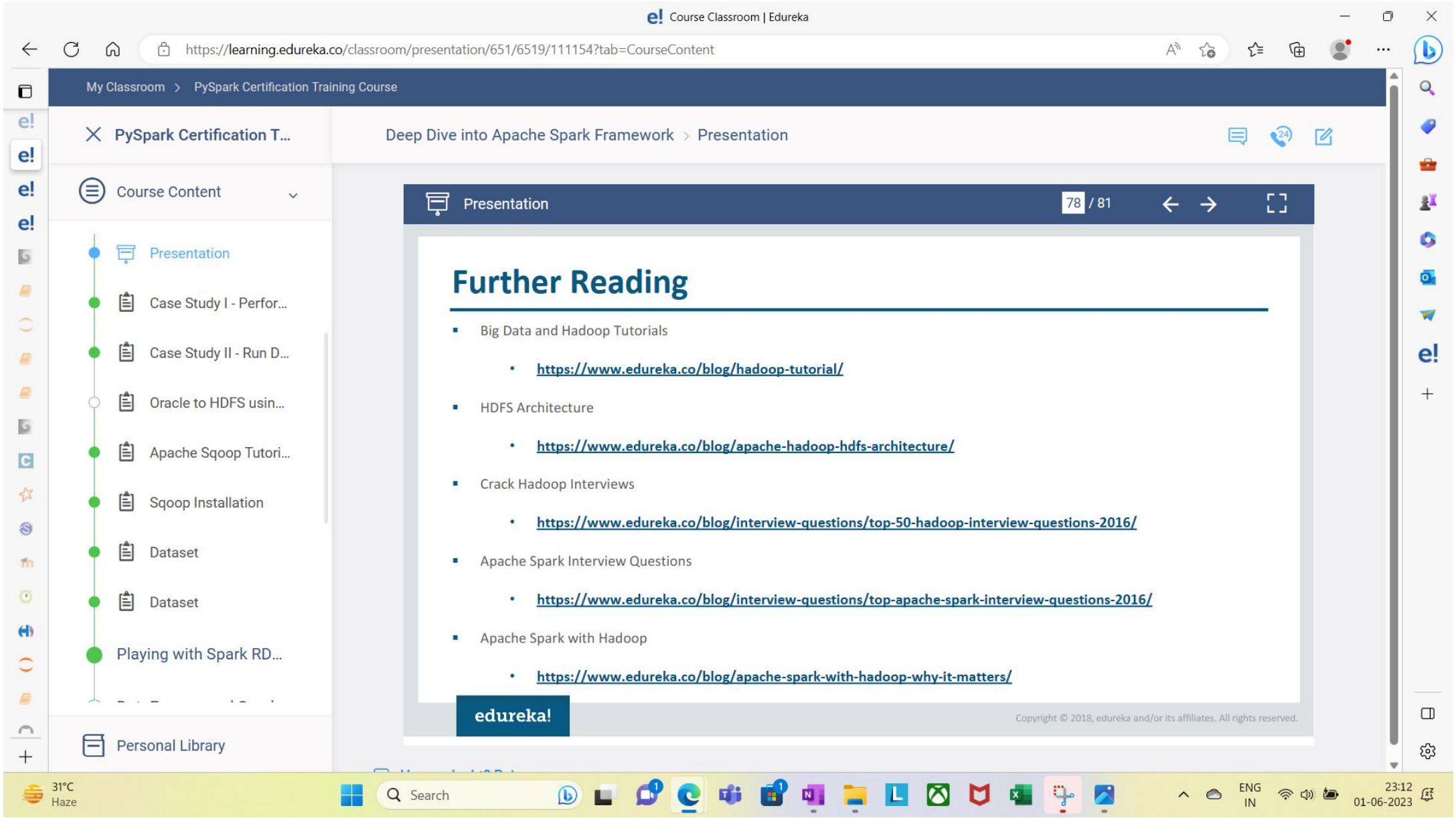






edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6519/111154?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Apache Spark Framework > Presentation
Course Content
Presentation
Case Study I - Perfor...
Case Study II - Run D...
Oracle to HDFS usin...
Apache Sqoop Tutori...
Sqoop Installation
Dataset
Dataset
Playing with Spark RD...
Personal Library
Presentation
78 / 81
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/
Apache Spark Interview Questions
https://www.edureka.co/blog/interview-questions/top-apache-spark-interview-questions-2016/
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.
31°C
Haze
Search
ENG
IN
23:12
01-06-2023