University of East London

Pioneering Futures Since 1898

Big Data Analytics - CN7031

Dr Amin Karami
Associate Professor in AI&DS

a.karami@uel.ac.uk
www.aminkarami.com

CN7031 – Lecture 1 (Intro)
Oct 2025

Outline

- The organization of Big Data module
- The tentative teaching topics
- The final assessment: A group-based CRWK
- The tools and software needed for Big Data

Module Aim

- This module aims to provide students with the core theoretical and practical background required for big data analytics and developing big data systems. It will provide you with an insight into areas of big data management and advanced analytics. You will develop in-depth practical skills through using tools and techniques from the forefront of the emerging field of data analytics.

Module Team

- **Lectures:**
  - Dr Amin Karami (**module leader**), a.karami@uel.ac.uk (office: EB.1.98)
  - Dr Fahimeh Jafari f.jafari@uel.ac.uk (office: EB.1.89)

- **Tutorial sessions:**
  - Dr Elias Eze, Dr Ali Jafari, Ms Dhara Parekh and 10 HPLs (PhD students and graduated MSc students)

Amin Karami

@AminKarami

Briefly speaking about my educational background, I completed my MSc D... >

aminkarami.com

- Please subscribe to the YouTube channel.
- This is essential for tracking your progress and contributions. We may follow up on attendance and participation for visa compliance if needed.

https://www.youtube.com/@AminKarami

Dr Fahimeh Jafari

- Senior Lecturer in the Department of Computer Science and Digital Technologies (CS&DT)
- Associate Director, Centre of FinTech
- Post Graduate Research Lead at CS&DT

f.jafari@uel.ac.uk

How to find CN7031 on Moodle?

https://moodle.uel.ac.uk/course/view.php?id=82303

CN7031 2526 (T1) Big Data Analytics (OC)

Amin Karami +1

View Course

Timetable

	Mon	Tue	Wed	Thu	Fri
08					
09	09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build - Docklands)	09:00 - 12:00 CN7031 [Big Data Analytics] RD.2.08 (RDCS Build -	09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build - Docklands)	10:00 - 12:00 CN7031 [Big Data Analytics] EB.2.44 - East Build	09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build -
10	09:00 - 12:00 CN7031 [Big Data Analytics] RD.2.25 (RDCS Build - Docklands)	09:00 - 12:00 CN7031 [Big Data Analytics] RD.2.08 (RDCS Build -	09:00 - 12:00 CN7031 [Big Data Analytics] RD.2.09 (RDCS Build - Docklands)	11:00 - 13:00 CN7021 [Advanced Software Engineering	09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build -
11	09:00 - 12:00 CN7031 [Big Data Analytics] RD.2.25 (RDCS Build - Docklands)	09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build -	09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build - Docklands)		09:00 - 12:00 CN7031 [Big Data Analytics] RD.1.15 - Compu
12		12:00 - 1 CN703 [Big Data Analyti RD.2.08 (RDCS Build -		12:00 - 13:00 CN7031 [Big Data	10:00 - 1 CN703 [Big Data Analyti RD.2.2 (RDCS Build -
13	13:00 - 1 CN703 [Big Data Analyti RD.2.09 (RDCS Build -	12:00 - 1 CN703 [Big Data Analyti RD.2.08 (RDCS Build -	13:00 - 16:00 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build - Docklands)	12:00 - 13:00 CN7031 [Big Data	10:00 - 1 CN703 [Big Data Analyti RD.2.2 (RDCS Build -
14	13:00 - 1 CN703 [Big Data Analyti RD.2.2 (RDCS Build -	13:00 - 1 CN703 [Big Data Analyti RD.2.0 (RDCS Build -	13:00 - 16:00 CN7031 [Big Data Analytics] RD.2.09 (RDCS Build - Docklands)	14:00 - 15:00 CN7031 [Big Data Analytics] MLT - Docklands Library	13:00 CN70 [Big Data Analy RD.1 (RDC Build
15	13:00 - 1 CN703 [Big Data Analyti RD.2.25 (RDCS Build -	13:00 - 1 CN703 [Big Data Analyti RD.2.2 (RDCS Build -	RD.1.18D (IT Lab) RDCS Build -		13:00 CN70 [Big Data Analy RD.2 (RDC Build
16	13:00 - 1 CN703 [Big Data Analyti RD.1.18 (RDCS Build -	14:00 - 1 CN703 [Big Data Analyti RD.1.1 (RDCS Build -	13:00 - 16:00 CN7031-SEMAD-12	16:00 - 17:00 CN7031 [Big Data Analytics] WB.G.02 - West Build (Docklands)	13:00 CN70 [Big Data Analy RD.1 (RDC Build
17	13:00 - 1 CN703 [Big Data Analyti RD.1.18 (RDCS Build -	16:00 - 19:00 CN7031 [Big Data Analytics] RD.2.09 (RDCS Build - Docklands) Group 35 - Ali	16:00 - CN70 [Big Data Analy RD.2.2 (RDCS Build	17:30 - 18:30 CN7031 [Big Data Analytics] ACE Online Session	13:00 CN70 [Big Data Analy RD.1 (RDC Build
18	13:00 - 1 CN703 [Big Data Analyti RD.2.25 (RDCS Build -	16:00 - 19:00 CN7031 [Big Data Analytics] RD.2.25 (RDCS Build - Docklands) RD.2.25E	16:00 - CN70 [Big Data Analy RD.1.1 (RDCS Build		14:00 CN70 [Big Data Analy RD.2 (RDC Build
			16:00 - CN70 [Big		14:00 CN70 [Big Data Analy RD.2 (RDC Build
					14:00 CN70 [Big Data Analy RD.2 (RDC Build
					16:00 - 19:0 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build - Docklands
					16:00 - 19:0 CN7031 [Big Data Analytics] RD.1.18 (RDCS Build - Docklands

---

- Timetabling: https://ueltt.uel.ac.uk

University of East London

Week 1

Lecture 1:
https://www.youtube.com/watch?v=9nj3UPRu2yk

Tutorial 1: See the Moodle Site

When the module starts:

Lecture: from 1st Oct onwards
Tutorial: from 3rd Oct onwards

Python: your ticket to awesomeness

Big Data is the future — and **Spark** + **Python (PySpark)** is your key to mastering it! 💾🔥

Why Python?

Easy to Learn: Python reads like English — no confusing symbols or crazy syntax. Perfect for beginners!
Versatile: From websites and AI to data processing and beyond, Python does it all!
Massive Community: Got stuck? Chances are, someone already solved it. The Python community is HUGE and super helpful.
Job Magnet: Python is the most in-demand language for careers in AI, data science, and software development.

Learn Python from zero

Part 1: Fundamentals
https://www.youtube.com/watch?v=93ldxMPtbq4&list=PL5d7U7T9Vj0qYg0D_42ac0zpRSDVKwmKH

Part 2: Visualization
https://www.youtube.com/watch?v=47GKEOI3qOM&list=PL5d7U7T9Vj0oWwjTkOnjejf8JopEbDyro

Module Assessment

- A **group-based** (3-5 members) CRWK (100%) including two tasks
- **Presentations**: week 12 (15th-19th December 2025)
- **Turnitin Submission**: 14th December 2025, 10pm
- In case of failure or non-submission of CRWK, there's no need to worry. You can have a 2nd attempt, but the highest grade you can get is capped to 50%. This usually takes place in March or April.
- All the group members must attend the presentation. If you do not attend, you fail the module. Every member of a group will be assessed individually.

Tentative Module Contents

- **W 1 (1-2 Oct):** Module Intro on Big Data [Amin]
- **W 2 (8-9 Oct):** Hadoop + HDFS [Amin]
- **W 3 (15-16 Oct):** Data Acquisition using Sqoop [Fahimeh]
- **W 4 (22-23 Oct):** Hadoop Hive for Structured Data [Fahimeh]
- **W 5 (29-30 Oct):** Unstructured Data in Hadoop [Amin]
- **W 6 (5-6 Nov):** Apache Spark + Spark DF [Fahimeh] [CRWK handout]
- **W 7 (12-13 Nov):** Spark Data Source API + Spark RDD I [Fahimeh]
- **W 8 (19-20 Nov):** HuggingFace with PySpark [Amin]
- **W 9 (26-27 Nov):** Spark RDD II [Amin]
- **W 10 (3-4 Dec):** Spark RDD for Unstructured Data [Amin]
- **W 11 (10-11 Dec):** Module review + CRWK Preparation [Amin]
- W 12 (15^{th} – 19^{th} Dec.): CRWK Presentation and Submission

After Completing Big Data: Career

CCA Spark and Hadoop Developer Exam (CCA175)

- **Number of Questions:** 8–12 performance-based (hands-on) tasks on Cloudera Enterprise cluster. See below for full cluster configuration
- **Time Limit:** 120 minutes
- **Passing Score:** 70%
- **Language:** English

https://www.cloudera.com/services-and-support/training/cdhhdp-certification/cca-spark.html

---

University of East London

Other Certifications

Cloudera Certified Professional (CCP)

Certification overview	Description
CCP Data Engineer	CCP Data Engineers possesses the skills to develop reliable, autonomous, scalable data pipelines that result in optimized data sets for a variety of workloads.

Cloudera Certified Associate (CCA)

Certification overview	Description
CCA Spark and Hadoop Developer	A CCA Spark and Hadoop Developer has proven their core skills to ingest, transform, and process data using Apache Spark™ and core Cloudera Enterprise tools.
CCA Data Analyst	A CCA Data Analyst has proven their core analyst skills to load, transform, and model Hadoop data in order to define relationships and extract meaningful results from the raw input.
CCA Administrator	Individuals who earn the CCA Administrator certification have demonstrated the core systems and cluster administrator skills sought by companies and organizations deploying Cloudera in the enterprise.
CCA HDP Administrator Exam	The HDP Certified Administrator (HDPCA) exam has five main categories of tasks that involve: Installation, Configuration, Troubleshooting, High Availability and Security.

https://www.cloudera.com/services-and-support/training/cdhhdp-certification.html

Other Certifications

https://www.bigdataframework.org/big-data-certification/

Big Data Tools and Software **(on your personal PCs)**

- **Install VMWare (Trial or full version):**
  https://www.vmware.com/

- **Cloudera VMWare [5.5GB]**: Cloudera is a US-based software company that provides a software platform for all data engineering, such as big data.

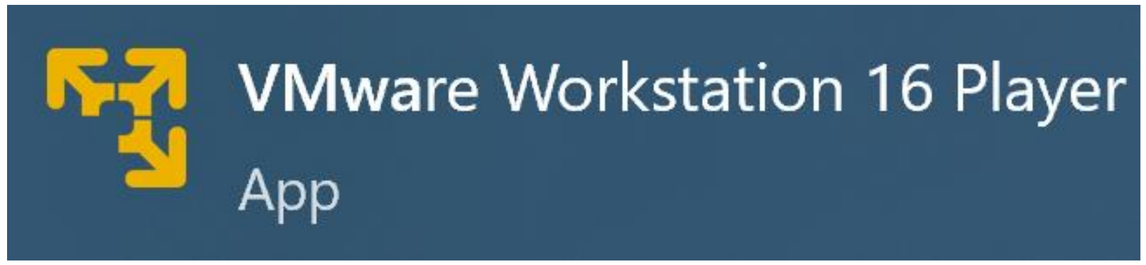
Weeks 2-5: https://tinyurl.com/699ck75s

Weeks 6-11: Google Colab: https://colab.research.google.com/

University of East London

Cloudera in the KD labs

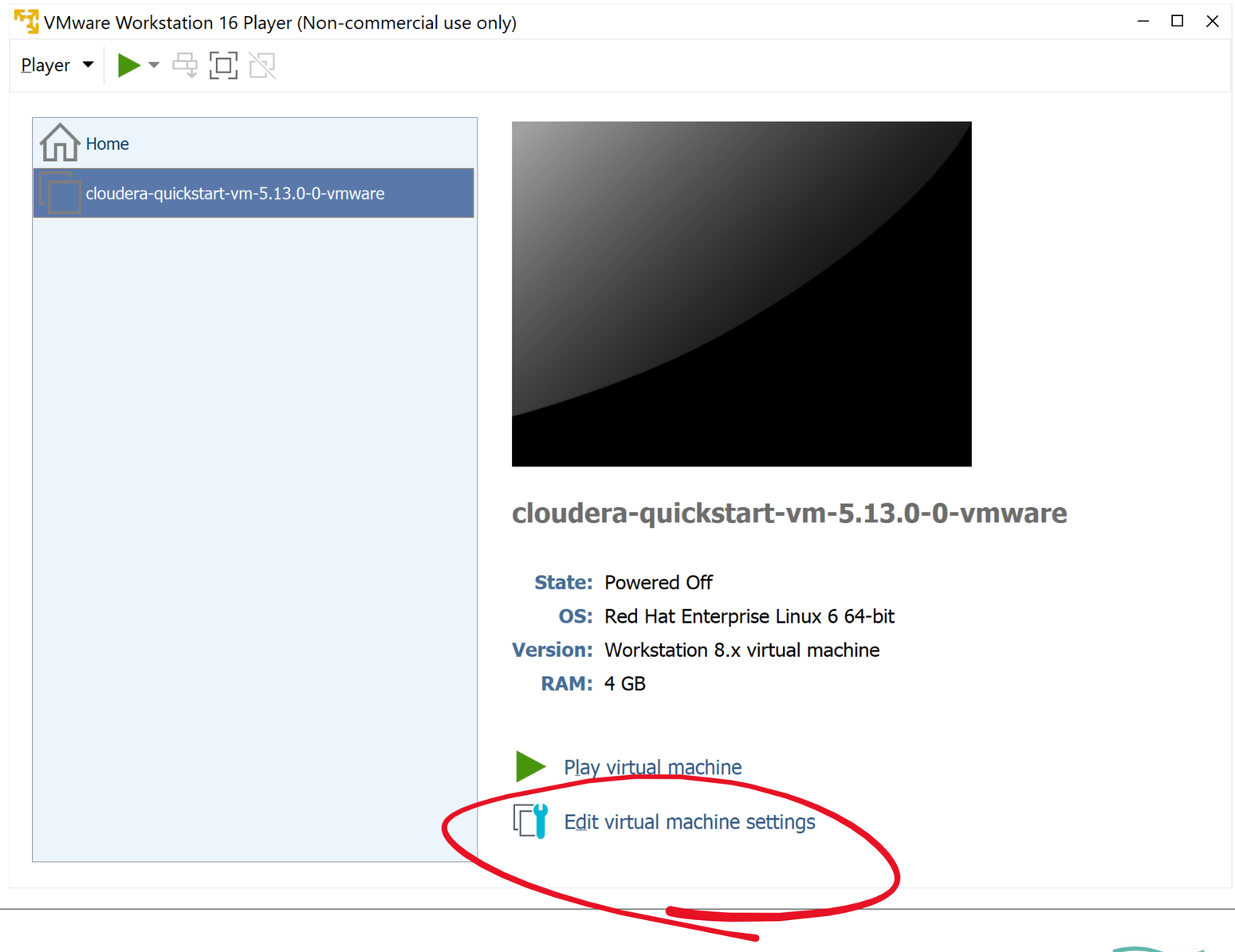
1- Open VMWare Workstation and then click on the “Open A Virtual Machine”



2- Locate “cloudera-quickstart-vm-5.13.0-0-vmware”, which is usually in Drive *C:/ACEVMData/cloudera 5.13*

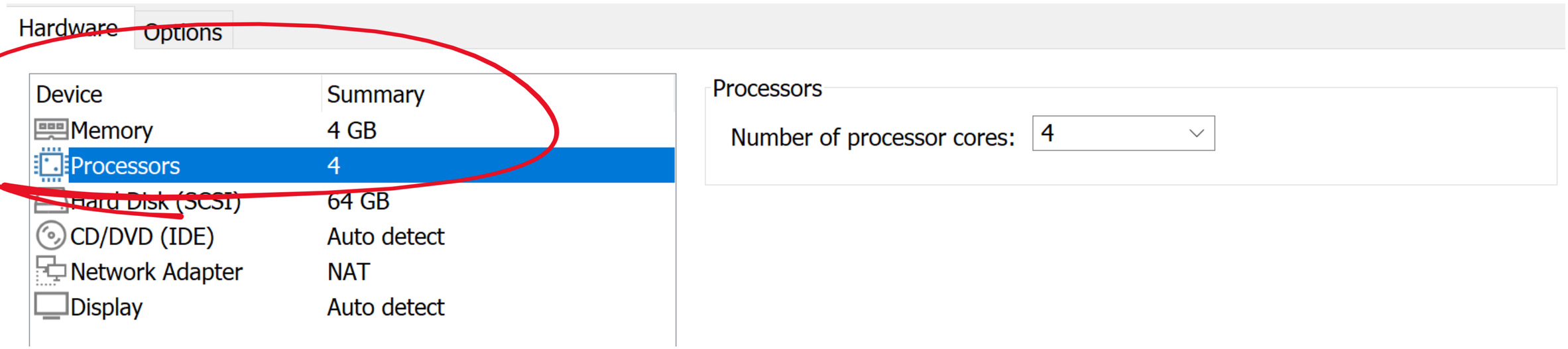
Cloudera in the KD labs

3- only for first time: click on “Edit virtual machine settings”



Cloudera in the KD labs

4- make sure to have at least 4GB RAM and 4 Processors



5- apply the settings and click on “Play virtual machine” to launch the VM. It takes ~5mins to launch the VM.

Cloudera in the KD labs

6- Click on “Launch Hue UI” or “Hue” on the toolbar

- *username:*

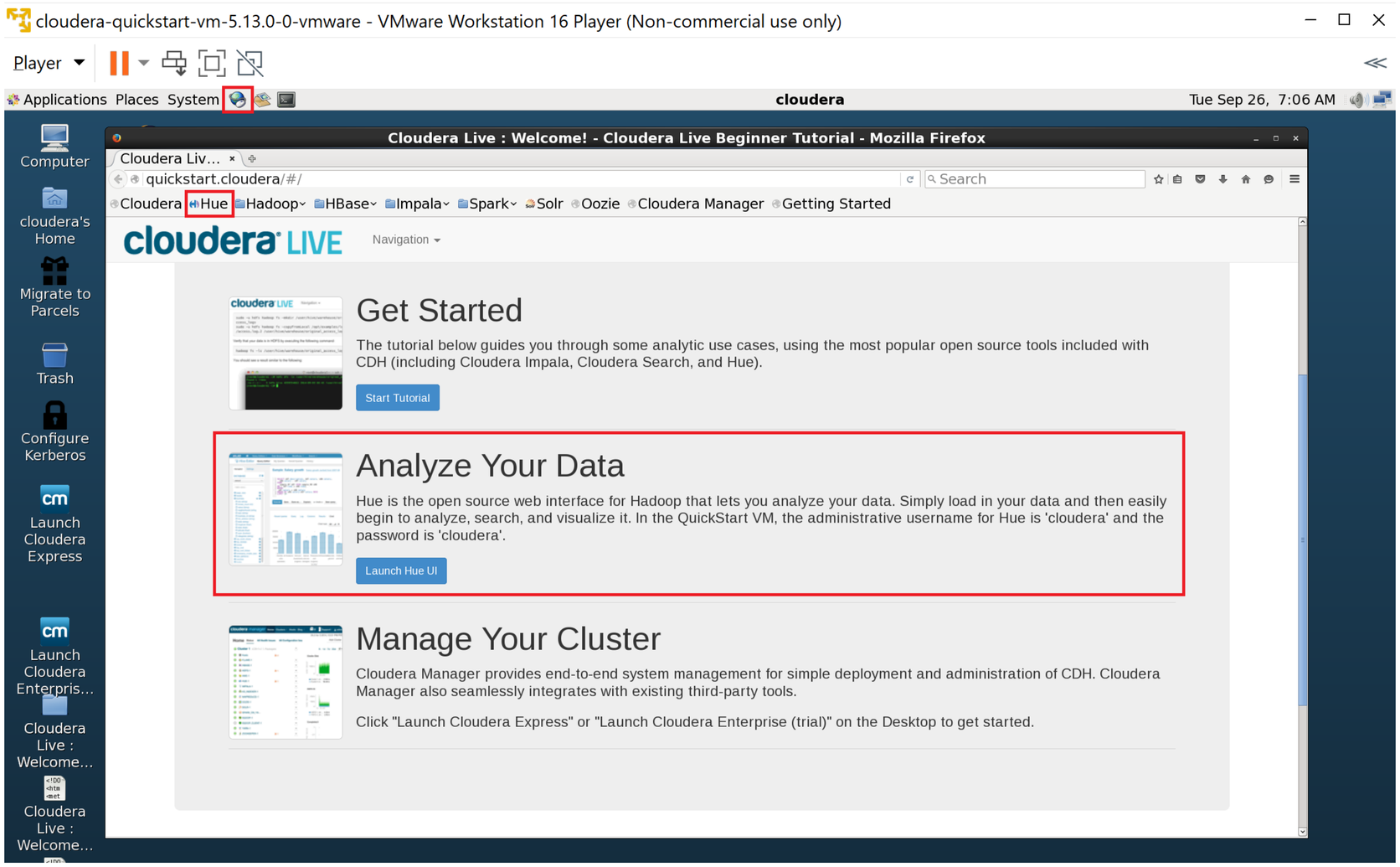
cloudera or **Big Data**

- *password*:

Cloudera

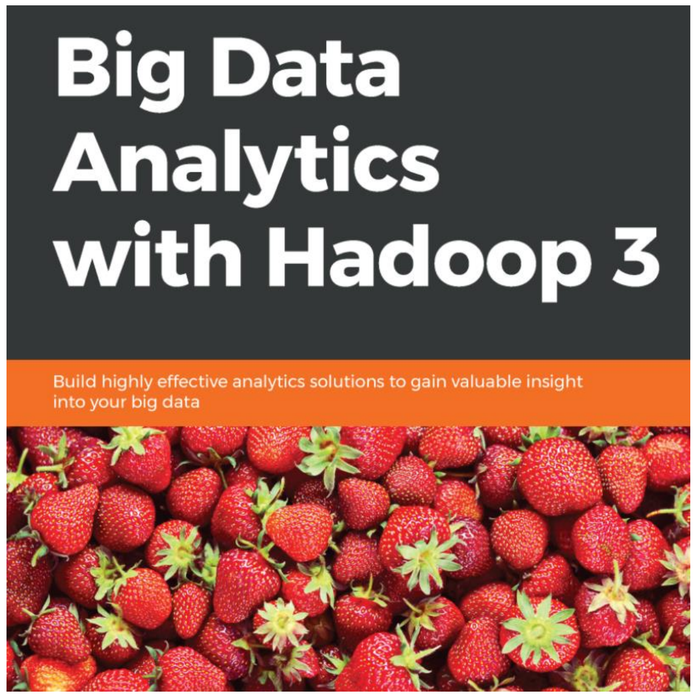
or

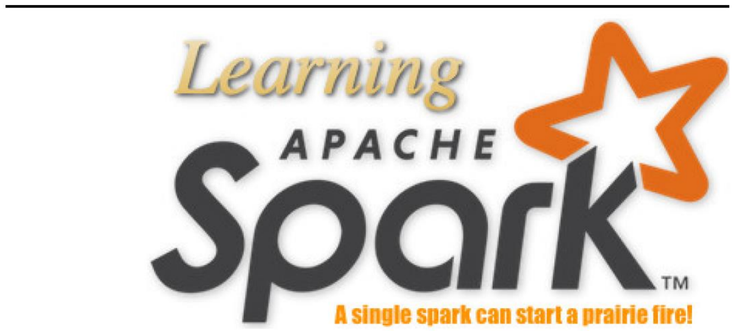
BigDataadmin

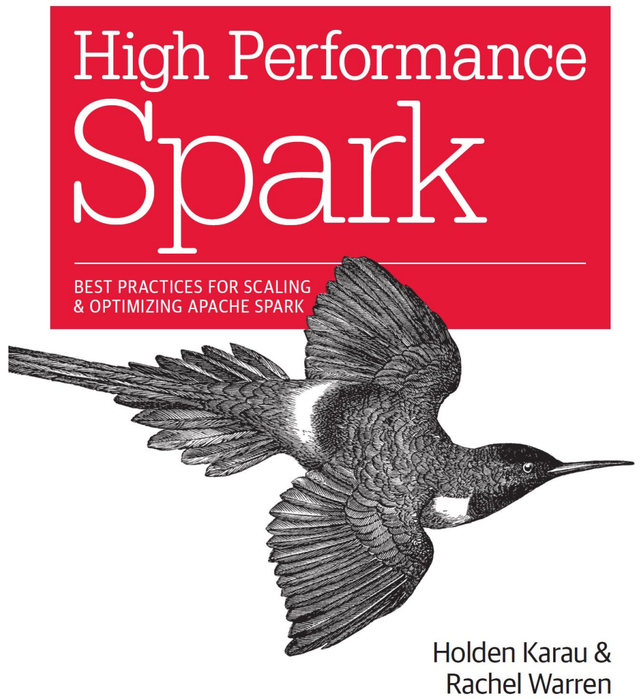


University of East London

Resources







https://tinyurl.com/y4w82ur9

https://tinyurl.com/yys7req9

https://tinyurl.com/yyw2ykss

University of East London

Summary

- The organization of Big Data module
- The tentative teaching topics
- The final assessment tips
- The tools and software needed for Big Data

University of East London