**UNIVERSITI TEKNOLOGI MARA**

**DSC761: ADVANCED DATA SCIENCE TECHNOLOGY**

| | |
|---|---|
| **Course Name (English)** | ADVANCED DATA SCIENCE TECHNOLOGY APPROVED |
| **Course Code** | DSC761 |
| **MQF Credit** | 3 |
| **Course Description** | This course will introduce students to the advanced data science technologies. The students will develop the skills of utilizing a set of tools as part of Big data ecosystem (e.g. Hadoop framework) for performing data extraction, cleaning and transformation. They will also develop their programming skills for data analysis, visualization and machine learning. The skills developed will then be used to solve appropriate case studies for data science with certain level of complexity. |
| **Transferable Skills** | Data extraction, cleaning, transformation, analysis, visualization and prediction using tools and programming language. |
| **Teaching Methodologies** | Lectures, Lab Work, Discussion |
| **CLO** | CLO1 Explain the advanced data science technologies<br>CLO2 Display the practical skills of utilizing the data science and engineering tools<br>CLO3 Experiment program in solving data science problems<br>CLO4 Demonstrate a prototype of a data science application |
| **Pre-Requisite Courses** | No course recommendations |

| **Topics** |
|---|
| **1. Fundamentals of Data Science**<br>1.1) Scientific Paradigm, Big Data Analytics, Data Challenges, Underlying Technologies, Reference Architecture, Big Data Techniques and Tools, Design Principles, Data Sciences with Business Mindset |
| **2. Hadoop Ecosystem**<br>2.1) Cloudera Hadoop, HDFS, MapReduce, YARN |
| **3. Database Management**<br>3.1) Data Definition and Data Manipulation (e.g. MySQL, Hive) |
| **4. Data Extraction**<br>4.1) Data Import and Export, Data Format Conversion, Data Compression, Data Extraction Job (e.g. Sqoop) |
| **5. Data Processing in Hadoop Ecosystem**<br>5.1) Data Cleaning and Transformation (e.g. Apache Spark based on Scala / Python Programming Languages) |
| **6. Data Processing and Analysis**<br>6.1) Python Programming (Data Structure, Functions, Flow Control, Matrix and Arrays) with Libraries (e.g. Pandas, Numpy) |
| **7. Data Visualization**<br>7.1) Data Preparation and Plotting using Libraries (e.g. Matplotlib, Seaborn, ggplot) |
| **8. Machine Learning**<br>8.1) Machine Learning Algorithms, Libraries for Machine Learning (e.g. Scikit-learn), Bias and Variance, Training and Testing, Validation Techniques |

| Assessment Breakdown | % |
|---|---|
| Continuous Assessment | 100.00% |

**Details of Continuous Assessment**

| Assessment Type | Assessment Description | % of Total Mark | CLO |
|---|---|---|---|
| Assignment | Assignment 1 (Part A and B) | 20% | CLO2 |
| Assignment | Assignment 2 (Part A and B) | 20% | CLO3 |
| Final Project | Can be group or individual project | 30% | CLO4 |
| Test | Final Assessment (open book test) | 30% | CLO1 |

**Reading List**

**Recommended Text**

- Jake VanderPlas 2016, *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media

**Reference Book Resources**

- Ofer Mendelevitch, Casey Stella and Douglas Eadline 2016, *Data Science with Hadoop*, Addison-Wesley Data & Analytics
- Garrett Grolemund and Hadley Wickham 2016, *R for Data Science*, 1st Edition Ed., O'Reilly Media
- Michael R. Brzustowicz 2016, *Data Science with Java: Practical Methods for Scientists and Engineers*, 1st Edition Ed., O'Reilly Media
- Carl Shan, William Chen, Henry Wang, Max Song, 2015, *The Data Science Handbook: Advice and Insights from 25 Amazing Data Scientists*, The Data Science Bookshelf
- Cole Nussbaumer Knafli 2015, *Storytelling with Data: A Data Visualization Guide for Business Professionals*, 1st Edition Ed., Wiley

**Article/Paper List**

**Recommended Article/Paper Resources**

- Rao T, Mitra P, Bhatt R, Goswami A 2019, The big data system, components, tools, and technologies: a survey, *Knowledge and Information Systems*, 60 [ISSN: 02193116] http://10.1007/s10115-018-1248-0

**Reference Article/Paper Resources**

- Oussous, Ahmed and Benjelloun, Fatima Zahra 2018, Big Data technologies: A survey, *Journal of King Saud University - Computer and Information Sciences*, 30 [ISSN: 22131248] http://10.1016/j.jksuci.2017.06.001
- Zhang Q Yang, L Chen Z Li P 2018, A survey on deep learning for big data, *Information Fusion*, 42 [ISSN: 15662535] http://10.1016/j.inffus.2017.10.006
- Xu, Li Da and Duan, Lian 2019, Big data for cyber physical systems in industry 4.0: a survey, *Enterprise Information Systems*, 13 [ISSN: 17517583] http://10.1080/17517575.2018.1442934
- Daniel, Ben Kei 2019, Big Data and data science: A critical review of issues for educational research, *British Journal of Educational Technology*, 50, 101-1 [ISSN: 14678535] http://doi.wiley.com/10.1111/bjet.12595
- Jimenez-Marquez J, Gonzalez-Carrasco I, Lopez-Cuadrado J, Ruiz-Mezcua B 2019, Towards a big data framework for analyzing social media content, *International Journal of Information Management*, 44 [ISSN: 02684012] http://10.1016/j.ijinfomgt.2018.09.003

**Other References**

- Website *Cloudera Documentation* https://www.cloudera.com/developers/get- started-with-hadoop-tutorial.html
- Website *Hive Documentation* https://docs.cloudera.com/documentation/ enterprise/5-8-x/PDF/cloudera-hive.pdf
- Website *Sqoop Documentation* https://archive.cloudera.com/cdh5/cdh/5/ sqoop/SqoopUserGuide.html
- Youtube *Cloudera Quickstart VM* https://www.youtube.com/watch?v=HloGuAzP_H8

| | |
|---|---|
| | • **Youtube *Hive Tutorial***<br>**https://www.youtube.com/watch?v=tKNGB5lZ PFE** |