<table>
<tr><td>Course Title</td><td colspan="5"><b>Big Data Analytics</b></td></tr>
<tr><td>Course Code</td><td colspan="5"><b>DLWSS552</b></td></tr>
<tr><td>Course Type</td><td colspan="5"><b>Elective</b></td></tr>
<tr><td>Level</td><td colspan="5">Master (2nd Cycle) – Distance Learning</td></tr>
<tr><td>Year / Semester</td><td colspan="5">2 / 3</td></tr>
<tr><td>Teacher’s Name</td><td colspan="5"><b>Dr Christos Markides, Prof. Leonidas Anthopoulos</b></td></tr>
<tr><td>ECTS</td><td>10</td><td>Lectures / week</td><td>3</td><td>Laboratories / week</td><td>-</td></tr>
<tr><td>Course Purpose</td><td colspan="5">The term Big Data was coined in the early 2000s, where sciences like genomics, and nuclear physics first experienced the data explosion. Companies and organizations in the private and public sectors are capturing trillions of bytes of information. This information is the fruit of the information society, as data is collected from customers, suppliers, and operations, as well as millions of networked sensors embedded in the physical world. These devices are sensing, creating and communicating data from mobile phones and transportation systems, to CCTV systems and sensor networks. With a smartphone in every pocket, or a tablet in every purse posting information, and updating multimedia content over social media thus fuelling the volume of information generated at an even faster rate. To this end, data is no longer regarded as static or stale, that has fulfilled its purpose once it was collected. Data has become the raw material for business, and a tool to produce new form of economic value, yield innovation, and create new services.
The purpose of this course is to provide students with a holistic approach to Big Data, the data model for Big Data, and examine the nature and requirements of a Big Data components, as well as Big Data as a platform. The course will then introduce the Apache Hadoop architecture, the Hadoop file system for distributed data storage, and highlights of MapReduce, Spark, Pig, and Hive as a framework of Big Data technologies as well as other tools for distributed processing of large datasets, using simple programming models. The course will address the shift from traditional relational database design for cluster systems to large scale NoSQL distributed databases, and the current trends for NoSQL database, as well as the future of Big Data. The course is complimented with practical and real-world examples. Finally, the course will examine the Big Data Analytics Lifecycle and review basic data analytic methods using R, pythong and Tableau, as an integrated environment for data integration, analysis, calculation, and visualisation for obtaining insight.</td></tr>
<tr><td>Learning Outcomes</td><td colspan="5">Upon successful completion of this course, students should be able to:
• Assess the state of Big Data adoption across a number of industry sectors.
• Describe the Hadoop Architecture, and use the Hadoop file system and components.
• Define and describe the major characteristics of NoSQL databases.
• Deploy, manage, and interact with NoSQL databases using interfaces,</td></tr>
</table>

	programming languages, and queries. • Analyse structured, unstructured, and IoT data for obtaining insight. • Evaluate schemas for different types of data stores. • Apply key concepts of Data Analytics Lifecycle to tackle Big Data problems. • Develop technical skills to analyse data for data exploration (ingest, store and secure data).		
Prerequisites	None	Corequisites	None
Course Content	This module consists of the following chapters: • **Chapter 1** introduces the fundamentals of Big Data and the main concepts of Data Mining. • **Chapter 2** explains the architecture of the Hadoop Framework and foundation core components • **Chapter 3** examines the architecture and characteristics of NoSQL databases • **Chapters 4 and 5** introduces Big Data Clustering platforms as general-purpose frameworks for cluster computing for data science and applications • **Chapters 6 and 7** expand general-purpose frameworks with the applications of NoSQL database and disparage storage needs • **Chapter 8** introduces the theory and application of Data Analytics and the Data Analytics Lifecycle • **Chapters 9 and 10** discusses the theory and application of Data Analytics methods • **Chapter 11 and 12** discuses methods and develops skills for data pattern discovery, data exploration, analysis, visualisation, and communication		
Teaching Methodology	**Mode of Delivery: Distance Learning** The course is designed to introduce and explain the material students are expected to learn through an on-line learning environment. The on-line environment provides an opportunity for receiving on-line feedback from the Course Instructor during their study. In addition, students will be encouraged to interact both with other students and the instructor so as to feel part of an on-line community of learners that belong to the University network. The course content will be delivered through online material/notes, recorded lectures and/or narrated presentations. Therefore, students may be asked to download and study notes, tutorials and numerical exercises as well as watch recorded lectures/demonstrations or narrated presentations posted on the web, addressing the main concepts of a particular unit. Furthermore, the planned communication and the dynamic/online interaction activities between the course instructor and the students will include asynchronous communication tools (Discussion Forum) where students may be asked to participate, wherever appropriate, in an online forum posting their views on certain topics covered in a particular unit; and synchronous communication tools (instant messaging, such as Skype, chat rooms, video-conferencing, etc.), so that students may discuss on-line with the Instructor(s) and/or other students specific issues covered in a given unit.		

	In addition, a number of case study readings are also considered, in order to demonstrate the adoption, application, and practice of Big Data Analytics in various organizations and the industry. The case studies illustrate the characteristics of Big Data, as well as how Analytics help organizations to obtain insight.
Bibliography	The following textbooks are associated with topics considered at various points throughout this course. EMC Education Services (2015), ***Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data***, John Wiley & Sons, 1st ed. Krishnan K. (2013), ***Data Warehousing in the Age of Big Data***, Morgan Kaufmann, 1st ed. Ladley J. (2019), ***Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program***, Academic Press; 2nd edition. White T. (2015), ***Hadoop: The Definitive Guide***, O'Reilly Media, 4th ed. Sadalage P., Fowler M. (2012), ***NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence***, Addison-Wesley Professional, 1st ed. Lars G. (2011), ***HBase: The Definitive Guide***, O'Reilly Media, 1st ed. Capriolo E., Wampler D., Rutherglen J. (2012), ***Programming Hive***, O'Reilly Media, 1st ed. Anderson J. C., Lehnardt J., Slater N. (2010), ***CouchDB: The Definitive Guide***, O'Reilly Media. Available [Online]: http://guide.couchdb.org/ Grolemund G., Wickham H., (2016), ***R for Data Science: Import, Tidy, Transform, Visualize, and Model Data***, O'Reilly Media, 1st Edition. McKinney W. (2017), ***Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython***, O'Reilly Media, 2nd Edition. Meier M., Baldwin D., and Strachnyi K., (2021), ***Mastering Tableau 2021: Implement advanced business intelligence techniques and analytics with Tableau***, Packt Publishing, 3rd Edition. **An extensive reading list of relevant academic papers:** Sanjay G., Howard G., and Shun-Tak L., ***The Google File System***, (SOSP '03). ACM, New York, USA, October 2003. Jeffrey D. and Sanjay G., ***MapReduce: Simplified Data Processing on Large Clusters***, Commun. ACM 51, Jan. 2008. Chang F., Dean J. , Ghemawat S., Hsieh W. C., Wallach D. A., Burrows M., Chandra T., Fikes A., and Gruber R., ***Bigtable: A Distributed Storage System for Structured Data***. ACM Trans. Comput. Syst. 26, 2, Article 4, 2008.

	DeCandia G., Hastorun D., Jampani M., Kakulapati G., Lakshman A., Pilchin A., Sivasubramanian S., Vosshall P., and Vogels W., ***Dynamo: amazon's highly available key-value store***. In Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles (SOSP '07). ACM, New York, NY, USA, pp 205–220, 2007. Lakshman A. and Malik P., ***Cassandra: a decentralized structured storage system***. SIGOPS Oper. Syst. Rev. 44, 2, pp 35–40, April, 2010. Massa S., and TestaS., ***Data warehouse-in-practice: exploring the function of expectations in organizational outcomes***. Inf. Manage. 42, pp. 709-718, July, 2005. Wang J., and Liu B., ***Design of ETL Tool for Structured Data Based on Data Warehouse***. In Proceedings of the 4th International Conference on Computer Science and Application Engineering 2020, ACM, New York, NY, USA, Article 119, pp. 1–5, 2020. Arora R., Pahwa P., and Bansal S., ***Alliance Rules for Data Warehouse Cleansing***. In Proceedings of the 2009 International Conference on Signal Processing Systems (ICSPS '09). IEEE Computer Society, USA, pp. 743–747, 2009. Weber K., Otto B., and Österle H., ***One Size Does Not Fit All—A Contingency Approach to Data Governance***. J. Data and Information Quality 1, 1, Article 4, 27 pages, June, 2009. Gopalkrishnan V., Steier D., Lewis H., and Guszcza J., ***Big data, big business: bridging the gap***. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine '12). ACM, New York, NY, USA, pp. 7-11, 2012. Sinaeepourfard A., Garcia J., Masip-Bruin X., and Marín-Torder E., ***Towards a comprehensive data lifecycle model for big data environments***. In Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT '16). ACM, New York, NY, USA, pp. 100–106, 2016. Tsai, CW., Lai, CF., Chao, HC. et al. ***Big data analytics: a survey***. Journal of Big Data 2, 21, 2015.
Assessment	The Students are assessed via continuous assessment throughout the duration of the Semester, which forms the Coursework grade and the final written exam. The coursework and the final exam grades are weighted 40% and 60%, respectively, and compose the final grade of the course. Various approaches are used for the continuous assessment of the students, such as dynamic online activities, online quizzes, group project design, implementation and presentation. The assessment weight, date and time of each type of continuous assessment is being set at the beginning of the semester via the course outline. An indicative weighted continuous assessment of the course is shown below: • **Three Online Quizzes** (15% of total marks for module)

	• **Two marked assignments** (20% of total marks for module) • **One dynamic interactive activity** (5% of total marks for module) • **One marked (group) project** (20% of total marks for module) • **One final written exam** (40% of total marks for module) Students are prepared for final exam, by revision on the matter taught, problem solving and concept testing and are also trained to be able to deal with time constrains and revision timetable. The criteria considered for the assessment of each type of the continuous assessment and the final exam of the course are: (i) the comprehension of the fundamental concepts and theory of each topic, (ii) the application of the theory in solving related problems and (iii) the ability to apply the above knowledge in complex real-life problems. The final assessment of the students is formative and summative and is assured to comply with the subject's expected learning outcomes and the quality of the course.
Language	**English**