

Course Title: Big Data & Analytics

Course Code: IS-628

Department: Information Systems

**HSM Vision**

HSM envisions its success in the sustainable contribution that it will make to the industry, academia and research in public and private sector. HSM will lead by providing professionally competent and ethically conscious human resources engaged in the global and local context to foster socio-economic growth and sustainability for the society. HSM envisages having faculty with high research potential and a deep desire for cutting edge research including collaboration with national and international partners.

**HSM Mission**

Being a research-oriented and student-centric business school, we emphasize research publications in impact journals as well as state-of -the-art learning methodologies.  We will prepare our students to become the future ethical business leaders and the guiding post for the society, while equipping them with the knowledge and skills required by world-class professionals.  We will be the leading choice for organizations seeking highly talented human resource. HSM will foster internationalization with key stakeholders and actively work to exchange best practices with business schools across Pakistan through collaborations, workshops, conferences and other means.

**Program Objectives**

The School of Business and Economics at UMT is foreseeing the challenges ahead both at national and international level and the utility of data science.  In Pakistan the multi-dimensional economy integrated with globalization needs a boost assisted by professionally trained and skilled Data Scientists, whom may incorporate and harmonize the unlimited bucket of resources, pouring in from springs of industry, agriculture, business, human resources etc. in a manner to achieve efficiency to its apex.  
In the competitive economy the companies need to adapt data science to gain a competitive advantage in productivity, profitability and sustainable business processes to offer better products and services to their customers. To attain this goal trained and skilled workforce in this area is the need of the hour; who are equipped to manage, understand and model the data, interpret the outcome and communicate the results for business use. Professionals holding a degree in Data Science will be well positioned to help their organizations gain a competitive advantage in a data-driven world.

**Course Objectives**

The Big Data course is designed to teach students the skills they need to work as Data Scientists. The skills we teach are the skills employers are seeking. It covers the spectrum of tools and techniques which are being adopted by businesses today to tackle data challenges, and the different roles that data specialists can fill in modern organizations. Students will gain an understanding of what insights big data can provide through hands-on experience with the tools and systems used by big data scientists and engineers.

Machine learning and IOT are undoubtedly one of the hot topics in data science right now, and this course aims to give a full overview, from theory to practical application using Big Data technologies. As well as an introduction to selecting data sources and tutored on selecting the best tools and frameworks for solving problems with data. Specialization will prepare for students to ask the right questions about data, communicate effectively with data scientists, and do basic exploration of large, complex datasets. Classes are a mixture of lectures, practical sessions at the computer and small group tutorials.

**Learning Objectives**

* They will be guided through the basics of using Hadoop with MapReduce, Spark and Hive.
* Discuss recent applications of Big Data technologies.
* Understanding of selecting data sources
* Understanding different frameworks used in Big Data as well as solving problems with data.

**Pre-requisites:**

* Knowledge of basic computer science principles and skills, at a level sufficient to
* Write a reasonably non-trivial computer program.
* Familiarity with the probability theory.
* Familiarity with linear algebra

**Teaching Methodology (List methodologies used –example are given below)**

* Lecture
* Interactive Classes
* Case based teaching
* Class activities
* Applied Projects
* Guest Lectures

**STUDENTS ARE REQUIRED TO READ AND UNDERSTAND ALL ITEMS OUTLINED IN THE PARTICIPANT HANDBOOK**

**Class Policy:-**

* Be On Time

You need to be at class at the assigned time. After 10 minutes past the assigned time, you will be marked absent.

* Mobile Policy

**TURN OFF YOUR MOBILE PHONE!** It is unprofessional to be texting or otherwise.

* Email Policy

**READ YOUR EMAILS!** You are responsible if you miss a deadline because you did not read your email.

Participants should regularly check their university emails accounts regularly and respond accordingly.

* Class Attendance Policy

A minimum of 80% attendance is required for a participant to be eligible to sit in the final examination. Being sick and going to weddingsare absences and will not be counted as present. You have the opportunity to use 6 absences out of 30 classes. Participants with less than 80% of attendance in a course will be given grade ‘F’ (Fail) and will not be allowed to take end term exams. International students who will be leaving for visa during semester should not use any days off except for visa trip. Otherwise they could reach short attendance.

* Withdraw Policy

Students may withdraw from a course till the end of the 12th week of the semester. Consequently, grade W will be awarded to the student which shall have no impact on the calculation of the GPA of the student. A Student withdrawing after the 12th week shall be automatically awarded “F” grade which shall count in the GPA.

* Moodle

UMT –LMS (Moodle) is an Open Source Course Management System (CMS), also known as a learning Management System (LMS). Participants should regularly visit the course website on MOODLE Course Management system, and fully benefit from its capabilities. If you are facing any problem using moodle, visit <http://oit.umt.edu.pk/moodle>. For further query send your queries to [moodle@umt.edu.pk](mailto:moodle@umt.edu.pk)

* Harassment Policy

Sexual or any other harassment is prohibited and is constituted as punishable offence. Sexual or any other harassment of any participant will not be tolerated. All actions categorized as sexual or any other harassment when done physically or verbally would also be considered as sexual harassment when done using electronic media such as computers, mobiles, internet, emails etc.

* Use of Unfair Means/Honesty Policy

Any participant found using unfair means or assisting another participant during a class test/quiz, assignments or examination would be liable to disciplinary action.

* Plagiarism Policy  
    
  All students are required to attach a “Turnitin” report on every assignment, big or small. Any student who attempts to bypass “Turnitin” will receive “F” grade which will count towards the CGPA. The participants submit the plagiarism report to the resource person with every assignment, report, project, thesis etc. If student attempts to cheat “Turnitin”, he/she will receive a second “F” that will count towards the CGPA. There are special rules on plagiarism for final reports etc. all outlined in your handbook.
* Communication of Results

The results of quizzes, midterms and assignments are communicated to the participants during the semester and answer books are returned to them. It is the responsibility of the course instructor to keep the participants informed about his/her progress during the semester. The course instructor will inform a participant at least one week before the final examination related to his or her performance in the course.

|  |  |
| --- | --- |
| Program | MS Data Sciences |
| Credit Hours | 3 |
| Duration | 15 Weeks |
| Prerequisites (If any) | N/A |
| Resource Person  Name and Email |  |
| Counseling Timing  (Room# 1N1 R#7 ) |  |
| Contact no. |  |
| Web Links:- (Face book, Linked In, Google Groups, Other platforms) |  |

**Chairman/Director Program signature………………………………….Date……………………..**

**Dean’s signature………………………… ………………….Date………………………………………….**

**Grade Evaluation Criteria**

Following is the criteria for the distribution of marks to evaluate final grade in a semester.

**Marks Evaluation Marks in percentage**

Quizzes 15%

Assignments 15%

Mid Term 25%

Attendance & Class Participation 5%

Term Project and Presentation 40%

Total 100%

**Reference Books:**

There is no required textbook for the class. There will be cases, articles as well as slides that we will distribute in each class.

**Optional Recommended Readings**

* For data mining and predictive analytics, the following are introductory textbooks: Data Science for Business, Provost and Fawcett: O’Reilly
* Data Mining for Business Intelligence, Concepts, Techniques and Applications, Shmueli, Patel, and Bruce: Wiley
* An interesting recent industry report on big data and the role that analytics may play in industry is:
* Big data: The next frontier for innovation, competition, and productivity and can be found at: http://www.mckinsey.com/insights/mgi/research/technology\_and\_innovation/big\_data\_the\_next\_frontier\_for\_innovation
* Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback Dataclysm: Who We Are (When We Think No One's Looking)
* Uncharted: Big Data as a Lens on Human Culture 1st Edition

**Software**

This course will require the use of hortonworks Hadoop.

**Course: -Machine Learning Course code: Book:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Topics to be**  **covered in the course** | **Learning Objective**  **of this topic** | **Expected Outcomes from Students** | **Teaching Method** | **Assessment Criteria** | **Deadlines and Homework** |
| 1 | Introduction to Big Data | Brief introduction toCourse: Big Data intro, Big Data Foundation, Big Data Applications and Big Data Trends | Students will be able to understand the basic terminology Big data | Lecture | Assignment | Within a Week |
| 2 | Deep understanding of Big Data | Understanding What's driving big data??Creating new value? Shifting power and influence, Managing the tradeoffs and Business Intelligence to Data Science | Students will be able to answer the advance level of Big Data | Case Discussion | Assignment | Within a Week |
| 3 | Data-driven Decision Making | Introduction to Data Analytics. Solving business problems using data analytics, Making business-defining decisions using data analytics, The data and analytics framework, Make better and faster decisions with data and analytics | They will be in position to understand the key values of Data-driven Decision Making | Case Discussion + Lecture | Class Activity | Within a Week |
| 4 | Technology and types of data | Learn the basics of Predicting the implications of technology advancements, "Structured", "Semi-Structured", and "Unstructured" data  Real-time Data Analytics: Data in action discussion and Identifying, organizing and processing data | After this lecture, students will be able to determine Technology and types of data | Lecture + Lab work | Quiz | Within a Week |
| 5 | Introduction to Hadoop | Understand what Hadoop is Learn about other open source software related to Hadoop. Understand how Big Data solutions can work on the Cloud, Hortonworks Introduction – HDP and HDF | They will be able to start working on Introduction to Hadoop | Case Discussion + Lecture | Class Activity | Within a Week |
| 6 | Hadoop Architecture | Understand the main Hadoop components, Learn how HDFS works, List data access patterns for which HDFS is designed, Describe how data is stored in an HDFS cluster | Students will understand Hadoop Architecture | Case Discussion + Lab Work | Quiz | Within two Weeks |
| 7 | Ethics in Hadoop Administration | Add and remove nodes from a cluster, Verify the health of a cluster, Start and stop a clusters component. Modify Hadoop configuration parameters and Setup a rack topology | Students will learn the Hadoop Administration | Lecture | Class Activity | Within two Weeks |
| 8 | Mid Term | Mid Term | Mid Term | Mid Term | Mid Term | Mid Term |
| 9 | Hadoop Components | Describe the MapReduce philosophy. how Pig and Hive can be used, how Flume and Sqoop can be used, Spark, Kafka and Nifi Introduction and how Oozie is used to schedule and control | After this lecture, students have understood various Hadoop Components | Case Discussion | Assignment | Within a Week |
| 10 | MapReduce and yarn | The MapReduce model v1. Review Java code to handle the Mapper class, the Reducer class, and the program driver to access MapReduce, The YARN model. Comparison of YARN / Hadoop 2 / MR2 vs Hadoop 1 / MR1 | Now they will be able to perform comparisons between MapReduce and yarn | Lecture + Case Study | Class Activity | Within a Week |
| 11 | Spark Fundamental | Introduction to Spark - Getting started, Components, Resilient Distributed Dataset (RDD), Spark standalone, Scala and Python overview, Launching and using the various Spark libraries | After this lecture, they are able to use Spark Fundamental | Lab Work + Case Study | Quiz | Within a Week |
| 12 | Resilient Distributed Dataset and DataFrames | Understand how to create parallelized collections and external datasets, Work with Resilient Distributed Dataset (RDD) operations, Utilize shared variables and key-value pairs | Their understanding with various Resilient Distributed Dataset and DataFrames | Lecture + Case Discussion | Assignment | Within two Weeks |
| 13 | Spark application programming | Understand the purpose and usage of the SparkContext, Initialize Spark with the various programming languages, Describe and run some Spark examples, Pass functions to Spark, Create and run a Spark standalone application and Submit applications to the cluster | They will be able to understand Spark application programming | Lab Work + Case Study | Quiz | Within two Weeks |
| 14 | Big data Performance Optimization | Spark configuration, monitoring and tuning, components of the Spark cluster, Configure Spark , environmental variables, Monitor Spark using the web UIs, metrics, and external instrumentation | Now they will be in position to understand Big data Performance Optimization | Lecture | Class Activity | Within a Week |
| 15 | Big Data and Machine Learning | Understand the difference between Dense and Sparse Data Types, and how they apply to Labeled Points and matrices, create and use the different matrices that are available in Spark MLlib, instantiate simple Linear Regression and Classification models, including Linear Regression, Support Vector Machines, and Logistic Regression. | Students will learn how the Big Data and Machine Learning will integrate. | Guest Speakers Presentation | QA session, Discussion |  |