BIG DATA TECHNOLOGY

Bachelor in Data and Business Analytics BDBA SEP-2023
BDT-DBA.3.M.A

Area Others
Number of sessions: 30
Academic year: 23-24
Degree course: THIRD
Number of credits: 6.0
Semester: 1º
Category: COMPULSORY
Language: English

Professor: **PEDRO FIGUEIRAS VICENTE**

E-mail: pfigueirasv@faculty.ie.edu

Senior Data Expert

- BSc Physics by USC. MBA by EEN Business school. Master's degree in AI research by UIMP.
- Executive education: Digital transformation by IE Business school. Internet of Things by MIOTI.
- 15+ years experience in Technology, both in Business and Engineering positions.
- C-level positions in Industrial companies.
- Freelance consultant for Data Engineering and Data Science projects.
- Private investor. Data & AI Startups mentor.

pfigueirasv@faculty.ie.edu

SUBJECT DESCRIPTION

Over the years, the value of information has increased, becoming a critical asset for companies, allowing them to make better decisions and increase the revenue generated by data.

In addition, the market has been digitized in all its aspects, from transactions between B2B (Business to Business) companies, to the relationship with customers and users (B2C or business to Commerce). This has caused the digital footprint to increase exponentially. The new commercial channels are mainly digital, such as Apps (applications), Chats, or web pages. The information is no longer static and structured, but dynamic and multimedia, with multiple structured and unstructured formats.

While the digital revolution has been growing in the markets, the technology has been completely updated, multiplying its capabilities and reducing its costs. From monolithic and generally on-premise IT architectures based on commercial software, it has been converted to open-source software and cost-effective cloud environments. Endless possibilities and much less entry barriers, but also greater complexity, significant orchestration needs and massive amounts of data to be stored and properly managed.

Big Data systems come to respond to these needs, offering state-of-the-art technologies, innovative processes and several components to make the most of the data on this new dynamic and exciting scenario.

LEARNING OBJECTIVES

The objective of this course is to provide students with the knowledge to understand the main pieces that are necessary to build and operate a Big Data system. The regular flow of data in a generic business system will be studied, applying the specific technologies in each section, but also understanding the functional needs, and the theoretical risk points beyond the technology. Additionally, students will be provided with the economic vision to be able to build an adequate business case and guarantee the profitability of the entire process.

At the end of the course; students should be able to:

- Understand the different data sources, data types and roles around Big data
- Identify the necessary phases in data management to convert massive data into knowledge
- Know the current hardware and software components involved in a Big data system
- Decide which is the best technological approach and architecture to be applied in each case
- Use the right tools to be able to ingest, process, analyze and extract the information from a Big Data (batch) system.

TEACHING METHODOLOGY

Learning Activity	**Weighting**	**Estimated time a student should dedicate to prepare for and participate in**
Lectures	50.0 %	75.0 hours
Discussions	3.33 %	5.0 hours
Exercises in class, Asynchronous sessions, Field Work	6.67 %	10.0 hours
Group work	20.0 %	30.0 hours
Individual studying	20.0 %	30.0 hours
TOTAL	100.0 %	150.0 hours

PROGRAM

SESSION 1 (LIVE IN-PERSON)

Introduction to Big data I:

- Definition. The many V’s of Big data
- Professional roles

SESSION 2 (LIVE IN-PERSON)

Introduction to Big data II:

- The MapReduce model

- Hadoop
- Assignment IBD

SESSION 3 (LIVE IN-PERSON)

Introduction to Big data III:

- First generations of Big data
- Introducing Databricks CE

SESSION 4 (LIVE IN-PERSON)

Introduction to Big data IV:

- OLTP Vs. OLAP
- “As a Service” model. Cloud vendors

SESSION 5 (LIVE IN-PERSON)

Introduction to Big data V:

- Storage strategies. Data types
- In-class workshop: Databricks & SQL basic

SESSION 6 (LIVE IN-PERSON)

Introduction to Big data VI:

- Big data Economics. Concepts
- Big data projects. Factors

SESSION 7 (LIVE IN-PERSON)

Introduction to Big data VII:

- Data governance
- Data ethics

SESSION 8 (LIVE IN-PERSON)

Introduction to Big data VIII:

- Part 1 recap
- Learning by doing: Questionnaire

SESSION 9 (LIVE IN-PERSON)

Data Ingestion I:

- Data sources
- Storage alternatives. HDFS Vs Object

SESSION 10 (LIVE IN-PERSON)

Data Ingestion II:

- Extracting data. Tools
- Storage files extension

SESSION 11 (LIVE IN-PERSON)

Data Ingestion III:

- RDBMS Vs No-SQL
- ELT approach

SESSION 12 (LIVE IN-PERSON)

Data Ingestion IV:

- Data lakes
- In-class workshop: Loading data into the Cloud
- Assignment DI

SESSION 13 (LIVE IN-PERSON)

Midterm Exam

SESSION 14 (LIVE IN-PERSON)

Data Ingestion V:

- Data lakehouses
- Learning by doing: Delta lake

SESSION 15 (LIVE IN-PERSON)

Data Ingestion VI:

- Data warehouses
- 2nd Gen DWHs. Snowflake

SESSION 16 (LIVE IN-PERSON)

Data processing I:

- PySpark workshop: Spark architecture

SESSION 17 (LIVE IN-PERSON)

Data processing II:

- PySpark workshop: RDD transformations & actions
- Learning by doing: PySpark exercises

SESSION 18 (LIVE IN-PERSON)

Data processing III:

- PySpark workshop: Spark dataframes
- Learning by doing: Dataframes exercise

SESSION 19 (LIVE IN-PERSON)

Data processing IV:

- Data processing pipelines
- In-class workshop: Apache Beam (GCP > Dataflow)
- Assignment DP

SESSION 20 (LIVE IN-PERSON)

Data analytics I:

- The value layer
- ML applied to Big data

SESSION 21 (LIVE IN-PERSON)

Data analytics II:

- ML tools: Introducing MLflow modules
- In-class workshop: MLflow applied to LR regression model

SESSION 22 (LIVE IN-PERSON)

Data analytics III:

- BI strategy
- Visualization insights
- Assignment DA

SESSION 23 (LIVE IN-PERSON)

Data Management I:

- Data privacy
- Data regulations
- Privacy Engineering

SESSION 24 (LIVE IN-PERSON)

Data Management II:

- GDPR in depth
- Data security
- Applying Data givernance frameworks

SESSION 25 (LIVE IN-PERSON)

Data Management III:

- Data Fabric
- Data Mesh

SESSION 26 (LIVE IN-PERSON)

Group project – Introduction and project steps

SESSION 27 (LIVE IN-PERSON)

Group project – Validation of workstreams

SESSION 28 (LIVE IN-PERSON)

Part 2 Recap:

- Next steps in Big data technology
- Main concepts review

SESSION 29 (LIVE IN-PERSON)

Group project – Final presentation

SESSION 30 (LIVE IN-PERSON)

Final Exam

EVALUATION CRITERIA

Your final grade in the course will be based on both individual and group work of different characteristics that will be weighted in the following way:

criteria	percentage	Learning Objectives	Comments
Final Exam	20 %		
Intermediate Tests	20 %		
Individual Work	20 %		
Workgroups	20 %		
Class Participation	20 %		

RE-SIT / RE-TAKE POLICY

Class participation:

To be evaluated based on the following criteria:

- Quality (not quantity) of your participation in class discussion: The most important dimension of participation concerns what it is that you are saying. A high quality comment reveals depth of insight, rigorous use of case evidence, consistency of argument, and realism.
- Frequency refers to the attainment of a threshold quantity of contributions that is sufficient for making a reliable assessment of comment quality. The logic is simple: if contributions are too few, one cannot reliably assess the quality of your remarks. However, once threshold quantity has been achieved, simply increasing the number of times you talk does not automatically improve your evaluation. Beyond the threshold, it is the quality of your comments that must improve. In particular, one must be especially careful that in claiming more than a fair share of “airtime”, quality is not sacrificed for quantity.
- Finally, your attempts at participation should not be such that the instructor has to “go looking

for you”. You should be attempting to get into the debate on a regular basis.

You might want to avoid being classified as one of the following types of students:

- Repeaters, i.e., students that, consciously or unconsciously, make comments that are really just repeats/rephrasing of what has already been said (by other students, or you). This wastes time and adds nothing to learning.
- Ramblers, i.e., students that take a lot of time to say simple things or they may tell long personal/professional stories, or they roam into topics that are not relevant, or simply make low-quality comments just to participate. They waste valuable time and prevent other students from being able to participate.
- Students that have been distracted (by Facebook, etc.) or who have stopped paying attention and then, later on, when they realized they have missed a term or concept, they ask you about it.

Exams

There will be one partial exam and one final exam. For these exams, you must bring your own laptops (other electronic devices are not allowed). You are also allowed to bring up 2 one-sided A4 sheets paper for the exams as well as any applicable cheatsheet that was delivered during the course.

In order to pass the course, you need a minimum grade of 3.5 in the final exam. If your grade in the final exam does not reach the threshold value of 3.5, you will fail the course, even in the case in which your weighted average (computed using the table above) exceeds 5.0.

As per University Policy:

Each student has 4 chances to pass any given course distributed in two consecutive academic years (regular period and July period).

It is mandatory to attend 100% of the classes. Students who do not comply with at least 70% attendance will lose their 1st and 2nd chance, and go directly to the 3rd one (they will need to enroll again in this course the next academic year).

Grading for retakes will be subject to the following rules:

1. Those students who failed the subject in the first regular period will have to do a retake in July (except those not complying with attendance rules who are banned from this possibility).
2. Dates and location of the July retakes will be posted in advance and will not be changed. Please take this into consideration when planning your summer.
3. The maximum grade that a student may obtain in the 2nd exam session is 8 out of 10. Those students in the 3rd call will be required to attend 50% of the classes. If due to schedule overlap, a different option will be discussed with the professor in order to pass the subject.

BIBLIOGRAPHY

Recommended

- Martin Kleppmann. *Designing data-intensive applications.* ISBN 9781491903117 (Digital)
- Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira. *Hadoop Application Architectures.* O'Reilly Media. ISBN 9781491900079 (Digital)
- Bill Chambers, Matei Zaharia. *Spark: the Definitive Guide.* O'Reilly Media Inc.. ISBN 9781491912300 (Digital)

BEHAVIOR RULES

Please, check the University's Code of Conduct here. The Program Director may provide further indications.

ATTENDANCE POLICY

Please, check the University's Attendance Policy here. The Program Director may provide further indications.

ETHICAL POLICY

Please, check the University's Ethics Code here. The Program Director may provide further indications.