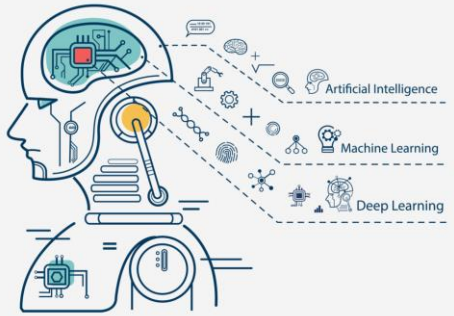

cct College Dublin
Computing • IT • Business

Big Data Storage & Processing

MSc in Data Analytics - Feb 2024 cohort

Module Introduction

CCT College Dublin

Ireland

Introduction

cct | College Dublin
Computing • IT • Business

- **Lecturer:** Dr. Muhammad Iqbal*
- **Experience:** Data Analytics, Numerical Modelling & Simulations, Structured & Object-Oriented Programming, Data Structures & Algorithms, Scalable Systems Programming (Python, R, Matlab, etc..).
- **E-mail:** miqbal@cct.ie
- **Contact:** Use CCT email address for contact along with your Studentid, Module and Course name.

Module Information

cct | College Dublin
Computing • IT • Business

- **Contact hours:**
  - 2.5 hours lecture and tutorials
  - More than 5 hours weekly approximately
- **Continuous Assessments**
  - 100% Continuous Assessment
- **Machine Requirements**
  - Windows 10/ 11 machine, At least 8 GB Ram and 512 GB hard disk, Core i7 or higher Micro processor

Objectives

cct | College Dublin
Computing • IT • Business

- The underlying concepts of Big Data Storage and Processing are mentioned below

1. Fundamentals of Big Data storage and data management paradigms
2. Underlying principles of parallel and distributed computing
3. Current solutions for retrieving, integrating and processing of Big Data
4. Big Data programming models and their efficient usage at scales
5. Big Data Streams and their processing techniques

Learning Outcomes

cct | College Dublin
Computing • IT • Business

- **On successful completion of this module, the learner will be able to**

1. Critically assess the **data storage and management requirements** of a given data project from a modern perspective and evaluate limitations of legacy approaches to Big Data. (Linked to PLO 3)

2. Assess the design concepts and architectural patterns of distributed Big Data systems and analyse the components that form their technology stack. (Linked to PLO 1, PLO 2)

3. Critically evaluate and select a Big data environment suitable for retrieving and processing a given Big Data set, perform data management and select appropriate analytic algorithms for the required scale and speed. (Linked to PLO 2, PLO 3)

4. Assess the functional differences between common Big data environments and particularities of Big data and appropriate Graph Big data and processing stacks. (Linked to PLO 4)

5. Implement the tools and technologies that facilitate the processing of Streaming Big Data to perform real-time analytics. (Linked to PLO 2)

Module Contents

cct College Dublin
Computing • IT • Business

Content
Legacy Approaches • Traditional Computing Architecture & Data Storage • Relational DBMS(SQL) & Data Silos • Old Data (SQL) vs. Big Data
Distributed Systems and Data Management • Architectures • Methodologies • Scaling
Big Data Storage • Physical Storage • Data Processing ETL/ELT • Data Tiering • File Formats, Compression and Security • Disaster Recovery
No-SQL • Key–value (e.g., Couchbase, Redis) • Document (e.g., MongoDB, CouchDB) • Columnar (e.g., Big Table, Cassandra) • Graph (e.g., Neo4j) • Spatial (e.g., OGC-compliant)

Big Data Platforms • Apache Hadoop and HDFS • MapReduce • YARN (resource management) • Apache Spark
Big Data Programming • Apache Hive (SQL-like queries) • Apache Pig (high-level scripts that run on Apache Hadoop) • Apache Mahout (machine learning algorithms on Apache Hadoop) • Spark MLlib (scalable and easy machine learning library on Apache Spark)
Streaming Big Data • Spark Streaming • Kafka
Graph Big Data • Apache Giraph (Graph processing on Graph Big Data)

CCT Resources

cct | College Dublin
Computing • IT • Business

- CCT ARC (https://arc.cct.ie/)
- For technical support, contact with Mr. Juan Murguey (jmurguey@cct.ie)

cct | College Dublin
Computing • IT • Business

Questions?

Books and eBooks

cct College Dublin
Computing • IT • Business

- Data Analysis with Python and PySpark, By Jonathan Rioux, Manning Publications, March 2022, 456 pages.
- Mastering Hadoop 3, Chanchal Singh, Manish Kumar, Packt Publishing, February 2019, 544 pages.
- Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, Thomas H. Davenport, Harvard Business Review Press, ISBN: 979-1422168165, 2014.
- Big Data: Concepts, Technology, and Architecture, Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry, Amir H. Gandomi, Mar 2021.
- Data Engineering with Apache Spark, Delta Lake, and Lakehouse, Manoj Kukreja, Packt, ISBN: 9781801077743, 2021.