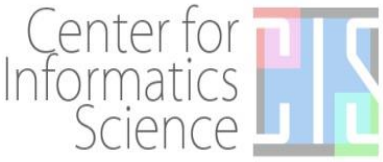

Big Data & Data Science Program

Diploma Courses

Date: Feb, 2016 – v 2.0

Diploma Structure

The Big Data & Data Science Diploma requires the attendance of 4 courses and 1 hands-on group project according to the following structure:

Semester #1 (2 Courses)

1- Introduction to Big Data, Developing with Spark and Hadoop (42 Hours, 14 Lectures)

2- Introduction to Machine Learning and Statistical Analysis (42 Hours, 14 Lectures)

Semester #2 (2 Courses + 1 Project)

3- Advanced Big Data Analytics Technologies and Applications (42 Hours, 14 Lectures)

4- Only 1 of the 2 following courses:

- Practical Data Mining (42 Hours, 14 Lecture)

OR

- Practical Data Science Using Machine Learning Technique (42 Hours, 14 Lectures)

5- Hands-on group project based on real life use case (6 Weeks of Mentoring)

Please refer to ***Appendix A*** for the description of each of those courses.

Important Notes

- All enrollments are subject to the admission rules and acceptance criteria of Nile University and the Big Data and Data Science Program.
- The default training location in Nile University premises and any change will be decided upon case by case by the program management team.
- Timing, lecture distribution, assigned instructors and schedules will be assigned and announced to students upon registration completion subject to Nile University and the program administrative decisions.
- The courses details and outlines might get changed due to continuous development and enhancements to cope with trending theories, technologies, methods and applications in this domain.

For more details and pricing, please contact us: bigdata@nu.edu.eg

Appendix A: Course Descriptions

CIT-650: Introduction to Big Data, Hadoop and Spark (42 Hours, 14 Lectures)

Description

The capability of collecting and storing huge amounts of versatile data necessitate the development and use of new techniques and methodologies for processing and analyzing big data. This course provides a comprehensive covering of a number of technologies that are at the foundation of the Big Data movement. The Hadoop architecture and ecosystem of tools will be of special focus to this course.

Students who complete this course will understand the architecture of Hadoop clusters at both the hardware and system software levels. Students will learn to apply Hadoop and related Big Data technologies such as MapReduce, Spark, Hive, Impala, and Pig in developing analytics and solving the types of problems faced by enterprises today.

Pre-requisites

- A course in operating systems
- Programming experience in Java, Python, or C/C++
- Recommended Backgrounds:
  - A general understanding of networking and distributed systems.
  - Familiarity with Linux and databases will be helpful.

(or equivalent knowledge subject to NU evaluation)

Reference Textbook

Main textbook: “Hadoop: The Definitive Guide,” (third edition); by Tom White

Other References:

“Hadoop Operations”, by Eric Sammer

“Programming Pig”, by Alan Gates

“Programming Hive”, by Capriolo, Wampler, and Rutherglen

Course Outlines

1. Introduction to Hadoop and MapReduce
   a. Hadoop Ecosystem
   b. Hadoop Clusters
   c. MapReduce API Concepts
   d. Basic Writing and testing MapReduce programs

2. Hadoop API
    a. ToolRunner Class
    b. HDFS programmatically
    c. Using the Hadoop API s Library of Mappers, Reducers and Practitioners
3. Managing Data Input and Output
4. Common MapReduce Algorithms
    a. Sorting and Searching Large Data Sets
    b. Indexing Data
    c. Computing Term Frequency
    d. Inverse Document Frequency (TF4IDF)
    e. Calculating Word Co4Occurrence

5. Joining Data Sets in MapReduce Jobs
6. Hadoop Tools for Data Acquisition
7. Practical Development Tips and Techniques
    a. Strategies for Debugging and Testing MapReduce Code
    b. Reusing Objects
    c. Creating Map4only MapReduce Jobs
8. PIG
    a. Complex Data Analysis with Pig
    b. Multi Dataset Operations with Pig
    c. Extending Pig
    d. Pig Troubleshooting and Optimization
9. Hive
    a. Relational Data Analysis with Hive
    b. Hive Data Management
    c. Text Processing with Hive
    d. Hive Optimization
    e. Extending Hive
10. Analyzing Data with Impala
11. Introduction to Spark
    a. Spark Basics
    b. Working with Resilient Distributed Datasets (RDDs)