Course program and reading list

Semester 2 Year 2021

School: Efi Arazi School of Computer Science M.Sc.

Big Data Platform

Lecturer:

Dr. Shaul Dar shaul.dar@post.idc.ac.il

Teaching Assistant:

Mr. Zuriel Levi zuriel.levi1@post.idc.ac.il

Course No.:	**Course Type :**	**Weekly Hours :**	**Credit:**
3605	Lecture	3	3

Course Requirements :	**Group Code :**	**Language:**
Final Paper	212360501	English

Prerequisites

Prerequisite:

52 - Calculus I
53 - Calculus II
54 - Linear Algebra I
55 - Linear Algebra II
56 - Discrete Mathematics
59 - Data Structures
69 - Logic And Set Theory
417 - Introduction To Computer Science

Course Description

Week	Topic
1	Review of key RDBMS concepts (e.g. ACID properties, SQL, query optimization) Assignment 1: SQL
2	Big Data: definition, motivation, use cases
3	The history of distributed databases, shared disk vs. shared nothing, Big Table, Hadoop, HDFS, Columnar DBs, Map/Reduce
4	Assignment 2: Map/Reduce
5	NoSQL DBs: concepts. CAP theorem, consistency models. Classification: key-value, wide column, document, graph
6	Review HBase, Cassandra, DynamoDB. Optional: High level overview of MongoDB or CouchDB, Neo4j or Amazon Neptune.
7	Assignment 3: Working with NoSQL DB
8	Message queues, Kafka
9	Spark, PySpark, DataFrames, Spark Streaming, SparkSQL
10	Assignment 4: read and process streaming data with Kafka and Spark Streaming
11	Spark ML (MLlib, spark.ml), Spark ML code walk thru

Course Goals

Big Data in combination with Machine Learning are transforming the world around us. Learn the key concepts and programming paradigms in this exciting technological frontier, including Hadoop, NoSQL Databases, and Spark, and how they can be used to store, process and analyze massive data sets. In the final project you will put all the pieces together in order to collect data and build a scalable machine learning model with Spark MLLib.

This course uses Python and a practical hands-on approach that will teach you how to approach a broad set of large scale data analysis problems effectively.

12	Final Project Practice! Including data ingestion, preprocessing, training and validating a model, evaluation.
13	Final Project Presentation.

Grading

There will be 4 assignments and a final project. All of them are mandatory.

Final grade = 0.4*HW + 0.6*Final project.

HW assignments and final project will be done in pairs. Only exceptional cases will be considered for individual submission.

Learning Outcomes

We will learn and apply the following technologies: Hadoop, HDFS, Map/Reduce, NoSQL Databases, Kafka, Spark (PySpark, DataFrames, Spark Streaming, SparkSQL) and Spark ML.

Lecturer Office Hours

Thursday, exact time TBD.

Tutor Office Hours

TBD

Teaching Assistant

Mr. Zuriel Levi, zuriel.levi1@post.idc.ac.il

Reading List

1. Practical Big Data Analytics: Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R (though IMHO you can skip the R section)

2. Hadoop The Definitive Guide