Course program and reading list

Semester 1 Year 2022

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

Big Data Platform

Lecturer:

Dr. Gil Vernik gil.vernik@post.idc.ac.il

Teaching Assistant:

Mr. Sahar Millis sahar.millis@post.idc.ac.il

Dr. Gil Vernik gil.vernik@post.idc.ac.il

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

Course Requirements :	Group Code :	Language:
Final Paper	221360501	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

We will cover major aspects of the Big Data trend focusing both on the popular Big Data engines for data processing and storage platforms for Big Data. Starting with basic definitions and notations like fault tolerance, consistency, CAP theorem, etc. We will deep dive into MapReduce paradigm, learn how Apache Spark internally works and how it differs from other Big Data engines, like Apache Hadoop MapReduce. We will learn various storage platforms that capable to persist large amount of data, like HDFS, columnar storage types like Apache Parquet, and cloud object storage exposing REST API. We also cover new emerging technologies like serverless computing and Function as a Service which recently attracts both academia and industry. We will define serverless and learn different serverless platforms, like Cloud Functions, Apache OpenWhisk, etc. and the benefit they provide for Big Data processing. We will cover various popular open-source projects, understanding theory and motivations through practical exercises and experiments.

Course Goals

1. **Introduction. Basic Terminology and definitions.**

Consistency, fault tolerance, replication, CAP theorem, etc.

2. **Data objects, columnar vs row storage, Apache Parquet, REST protocol, metadata, etc.**
3. **Storage for Big Data. Distributed file systems. HDFS. Object Storage. Difference between object storage and HDFS. Use cases and challenges.**
4. **MapReduce. Understanding the MapReduce concept, shuffle, fault tolerance aspects, data locality, partitions discovery.**
5. **Big Data processing engines. Apache Hadoop MR and Apache Spark. Resilient Distributed Datasets, etc.**
6. **Serverless and Function as a Service. Containers and VMs. Understanding the concept, challenges, use cases and Big Data processing with Serverless.**

Grading

Home assignments - 20%

Final project - 80%

2 points bonus for students who attended lectures (attendance is maintained via EZCheck.me!)

Lecturer Office Hours

Tuesday, 20:00-21:00, via Zoom

Reading List

1. https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
2. https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
3. https://arxiv.org/abs/1709.01812
4. https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf
5. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36632.pdf
6. http://elmeleegy.com/khaled/papers/delay_scheduling.pdf
7. https://dl.acm.org/citation.cfm?id=3284029
8. https://dl.acm.org/doi/abs/10.1145/3429880.3430101