Databases & Big Data Analysis

Author: Dr. Keren Ouaknine

This elective course is open for second, third year or master students (no previous db knowledge).
Duration: 3 academic hours per week (3 x 45mn) for 13 weeks.

List of topics

- Query data using **SQL**.
- Launch a **Hadoop** cluster on Amazon, and querying data using **Pig and Hive**.
- Launch a **Spark** cluster on Amazon (or GCP), and querying data using Python/Scala/Java
- Mining **streams** of Data using **Kafka** concatenated to Storm
- Learn about the **architecture of Big Data platforms**: components, and use cases.
- Query data from a notebook on top of Hadoop/Spark using **Zeppelin, Hue**
- Learn **visualization tools** with Tableau, etc.
- Get familiar with **Graph databases** and social querying.
- Differentiate between **key-value** stores, **columnar** stores, and **document** stores.
- Two weeks on ML topics: intro to ML, deep learning and TensorFlow
- Develop a **final end-to-end project**: from raw data to BI insights.

Optional topics: Impala, Drill, Phoenix, etc

Tentative Timeline:

Week 1: Oct 24

- SQL Operators part 1: selection, distinct, union, aggregation, join and their types: left-outer join, inner join. Join implementations: replicated join, nested loop join, hash join, nested queries.

- Review of the ACID and CAP theorem and their applications in the industry.
Mention WAL (Write-ahead logging provides “all or nothing” failure atomicity, database, consistency, and transactional durability).

- Indices: B+ tree, R-Tree.

Complementary pointers to read:

→ chapter 5.2 select, 5.3 union, 5.5 aggregation, 5.6.4 joins in the Ramak book
→ chapter 8 -8.3 (index, B+ tree) in Ramak book.
→ chapter 5.4 Ramak for nested queries

Week 2: Nov 1

- MapReduce programming model, Hadoop and HDFS. replication (network topology), data shuffling, and task speculation.

Homework - reading:
- Simplified data processing on large clusters by Dean, and Ghemawat (aka as the Google white paper on MapReduce)
- Designing good MapReduce algorithms by Prof. J. Ullman

Week 3: Nov 7

- Query languages on top of Hadoop: Pig (procedural), Hive (declarative), Presto.
- Query optimizers. Rule based, cost based optimizer.

Homework:

Hands-on exercise on programming with Pig and Hive.

+ paper exercise on query optimization.

Week 4: Nov 14

- Spark architecture and use cases.

Homework:

Use Spark to process data - part 1

Week 5: Nov 21 + guest lecturer from myHeritage on Kafka

- Mining data streams with Kafka and Storm on top of Spark
- Agile

Homework:

Build up a Kafka pipeline and use Spark to process data - part 2

Week 6: Nov 28

- Notebooks: Zeppelin, HUE on top of Hadoop.
- Visualization tools: Tableau, Power BI, and Zoho on top of EMR

Homework: Hands-on exercise using Zeppelin

Week 7: Dec 5 + guest lecturer by Gilad Engel on the startup eco-system in IL/EU/US

Homework:

Project - part 1

Week 8: skip ~~Dec 12~~ (Boston) ⇒ **Dec 19**

- Graph databases: social networks, community detections.

Move into the NOSQL chapter (column, document & graphs, talk time-series & geospatial)

Homework:

Project - part 2

Week 9: Dec 26
- NoSQL databases: key-value stores, columnar stores (Apache Orc, Parquet, Kudu), document stores. Demo in class, use the Databricks notebook from Shay to demonstrate how Redis can extract information x40 faster. if it doesn't have to come from Spark. Once the learning is done in Spark, the 100k users are pre-built as a batch and loaded into Redis so that users can access it in submillisecond.
Homework:
Project - part 3

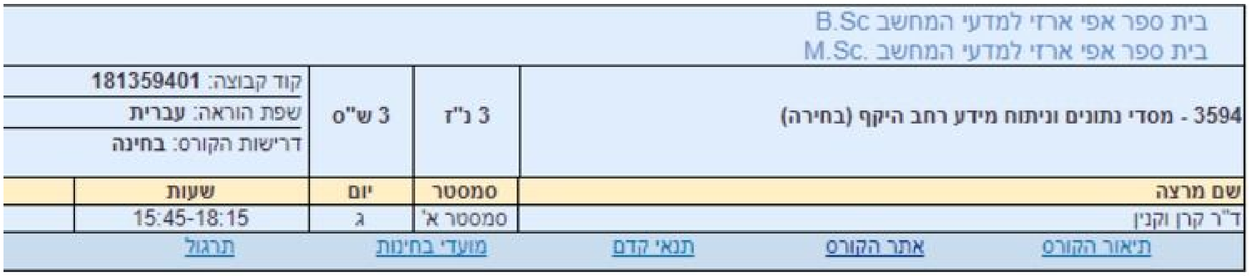
Week 10: ML topic 1- Jan 2 Machine Learning
Homework:
Project - part 4

Week 11: ML Topic 2 - Jan 9 Tensor Flow
Homework:
Project - part 5

Week 12: Presentation of final projects. Jan 16

Course grading:

- **Homework:** 50% of grade ⇒ **30% of grade**
- **Project:** 25% of grade ⇒ **35% of grade**
- **Final exam:** 25% of grade ⇒ **35% of grade**
-



בית ספר אפי ארזי למדעי המחשב B.Sc
בית ספר אפי ארזי למדעי המחשב M.Sc.

3594 - מסדי נתונים וניתוח מידע רחב היקף (בחירה)	3 נ"ז	3 ש"ס	קוד קבוצה: 181359401 שפת הוראה: עברית דרישות הקורס: בחינה	
שם מרצה	סמסטר	יום	שעות	
ד"ר קרן וקנין	סמסטר א'	ג	15:45-18:15	
תיאור הקורס	אתר הקורס	תנאי קדם	מועדי בחינות	תרגול

הקורס #3594 מתוכנן להכיל 4 תרגילים יבשים ו-5 תרגילים רטובים. בקורס יהיה גם פרויקט גדול וגם בחינה.

Textbook:
While we do not have any required textbook for the course, the following books will be useful references for the material that we will be covering in class.

1. A first course in database systems by J. Ullman and J. Widom
   Link here, waiting for library IDC referral (new book).

2. Database Management Systems, 3rd edition, by R. Ramakrishnan.

Library IDC referral: 005.74RAM Link here

Course schedule:
part A: 15:45-17:00
break: 17:00- 17:15
part B: 17:15 - 18:15
Reception hour: 14:30 - 15:30, by appointment only.

חנייה: משרד המנהלה - שרה בקומה 2
חניון בכניסה או חניון מערבי כניסה מרח' אלתרמן

Grade Change Policy: For all of the graded assignments as well as the final project and exam, if you disagree with the grading, you may discuss your concerns with the T.A. (Andrew Baronick) within **1 week** after they are returned. After that, all grades will be considered final.

Prerequisites

Since this is a hands-on class, you are expected to bring your laptop to every class (and remember to charge it, so that it lasts for the duration of the class).
+ setting up an Amazon Web Services account (AWS) and BigQuery too.

Course Policies

Unless, otherwise noted, we follow the default IDC Policies (add link).
Due to the hands-on nature of the course, absences are strongly and highly discouraged.
slides were created to support the lectures, not replace them.

Classes are mandatory from the very first week. If you registered late i.e within the first two weeks (a.k.a "change period") then you should attend these two weeks and submit your homework exercises on time so that you can register.
Please do not ask for grace period because you registered late, these will be automatically rejected.

Frequently Asked Questions

Q: I am not registered to class, but I really need this course. What should I do?
A: Attend the first two weeks, and submit your homework. If registered students cancelled their registration by the end of the "changing courses period" of two weeks from the semester starting date, you might be able to join. If you don't come to class or if you didn't submit the homework, you won't be able to join the class.

Q: Do I have to attend the classes?
A: Yes, most theoretical and hands-on content will take place during classes.

Q:

A:

Late Assignment Submission Policy
You are free to submit late, but there is a 3% grade penalty for every additional day after the deadline, and you can be at most 14 days late. Given the generous late submission policy, penalties are strictly enforced, and no extensions are granted. Please plan accordingly, and do not leave submission for the last minute.

Other option: clear 0.5 point for every hour late.

~~**Assignment late policy:** The official due date for each assignment will be listed on moodle, and it is expected that students will turn work in on time. We will offer a 48-hour "grace period" for each assignment, and we will accept solution submissions that come in within 48 hours of the due date (i.e., less than two days late). Late assignments will not be accepted beyond the grace period.~~

Homework:

5 wet and 4 dry exercises

1 ⇒ 5%

2, 3, 4 ⇒ 4 % each

5 ⇒ 1%

6 (viz) ⇒ 4%

3 quiz left : 3%, 3%, 2%

Optional:
- Datadog, Big Query, PySpark, and Presto, YARN, Tez, Impala,
- ask for an education grant for the course from Amazon.

Course description (for the shnaton):
The rise of Big Data and the vision of a data-driven world present many exciting new challenges related to processing Big Data, handling data diversity, exploiting new hardware, software, and cloud-based platforms.

This course provides an introduction to data processing using SQL and an overview of popular and modern platforms processing data (Hadoop, Spark, etc).

We begin with the relational model and the SQL language. We will study Big Data platforms, their processing methods, components and use-cases on the cloud. We will then discuss mining of streams, key-value stores, graph databases, and finally recommendation systems.

In the future, we could open up the course to non-CS:

This course is the recommended starting point for students who 1) are interested in jobs in the rapidly growing fields of data science and data analytics or 2) who are interested in acquiring the technical and data analysis skills that are becoming increasingly relevant in other disciplines such as finance and marketing.