



החוג לניהול / Dept. of Management

1221-7022 – טכנולוגיות ביג דאטה

## Big Data Technologies

דרישת קדם: טיפול יישומי בנתונים. תלמידי מדעי המחשב, מדעים להייטק פטורים מדרישה זו.

מסטר ב'- שנה"ל תשפ"ב

קבוצה	יום בשבוע	שעה	כיתה	מרצה	דואר אלקטרוני	טלפון
01	א' וגם ד'	10:00-12:00 * 12:00-14:00 *	404	ד"ר משה אונגר	mosheunger@tauex.tau.ac.il	03-6408112

- שעת קבלה – בתיאום מראש
- קורס חצי סמטריאלי – מחצית שנייה של הסמסטר, תאריכי מפגשים: 6.4.22 – 10.6.22

### היקף הלימודים

היקף הש"ס לקורס: 2 ש"ס

ECTS = 2 ש"ס = 4 ECTS (European Credit Transfer and Accumulation System), ערך הניקוד של הקורס במוסדות להשכלה גבוהה בעולם שהינם חלק מ"תהליך בולוניה".

### תיאור הקורס

מהפכת ה-Big Data משנה את העולם שבו פועלות חברות וארגונים, אין תעשייה שלא מושפעת מההשלכות של ניתוח נתוני עתק. על מנת להפיק תובנות אנליטיות בעלות ערך לארגון, נדרש לבצע פעולות קריטיות הכוללות איסוף, עיבוד, אחסון וניתוח נתונים במימדי ענק. לשם כך, נדרשת תמיכה במגוון של טכנולוגיות חדשניות הקרויות טכנולוגיות נתוני עתק. בקורס נלמד מהם הופך נתונים ל-Big Data, נכיר סוגי נתונים שונים – מובנים (Structured) ולא מובנים (unstructured) תוך התמקדות בתכנותיהם המאפשרות התאמה לניתוח נתוני עתק. וננתח את חשיבותם לעולם ה-Big Data, נלמד ונכיר את סכמת MapReduce וכן Data File System (DFS). בחלקו השני של הקורס נסקור טכנולוגיות שונות לפתרון הבעיות שנסקרו בחלקו הראשון: Hadoop, Spark, AirFlow, Kafka. הקורס יספק בסיסי תיאורטי שיאפשר העמקה בתחום וכן התנסות מעשית בטכנולוגיות נתוני עתק. בסיס הקורס נדבר על השלכות של Big Data בדגש על רגולציה – GDPR וכן היבטים עסקיים ואתים אחרים. יהיו שני מפגשי תרגול במהלך הסמסטר בהן נתרגל מימוש של טכנולוגיות נתוני עתק ב-Spark.

## תפוקות למידה

עם סיום הקורס בהצלחה יוכלו הסטודנטים:

1. להבין מהם האתגרים העומדים בפני גופים בעידן Big-Data
2. לתאר מהי סביבת Hadoop
3. לדעת כיצד ניתן לממש אלגוריתמים נבחרים בסביבת MapReduce
4. לעבוד ברמה בסיסית עם Spark

## הערכת הסטודנט בקורס והרכב הציון

אחוז	מטלה	תאריך	גודל קבוצה/ הערות
5%	השתתפות ותרומה לדיון		
20%	תרגילי כיתה		הגשה ביחידים
75%	מבחן מסכם		חובת מעבר בציון של 60 לפחות

הערות לגבי ציונים:

- תלמיד חייב להיות נוכח בכל השיעורים.
- תלמיד הנעדר משיעור המחייב השתתפות פעילה או שלא השתתף באורח פעיל, רשאי המורה להודיע למזכירות כי יש למחוק את שמו מרשימת המשתתפים (התלמיד יחויב בתשלום בגין קורס זה).

## מדיניות שמירה על טווח ציונים

בחוג לניהול מונהגת מדיניות שמירה על טווח ציונים. מדיניות זו מתייחסת לממוצע הציונים הסופיים בקורס. מידע בנושא זה מתפרסם בהרחבה באתר החוג לניהול, בסעיף ציונים בתקנון.

## הערכת הקורס ע"י הסטודנטים

בסיומו של הקורס הסטודנטים ישתתפו בסקר הוראה על מנת להסיק מסקנות לטובת צרכי הסטודנטים והאוניברסיטה.

## אתר הקורס

אתר הקורס יהווה המקום המרכזי בו ימסרו הודעות לסטודנטים, לפיכך מומלץ להתעדכן בו מדי שבוע, לפני השיעור, ובכלל – גם בתום הסמסטר (לצורך תיאום ענייני הבחינה למשל).

שקפי הקורס והתרגול יועלו לאתר הקורס באתר.

לתשומת לבכם - בהרצאות ידונו גם נושאים (ובפרט דוגמאות) שאינם מופיעים בשקפים או מופיעים בכותרת בלבד. כל אלו הינם חלק בלתי נפרד מחומר הקורס.

## תכנית הקורס \*

Topic Number	Topic	Optional Reading	Comments
1	<b>Introduction to Big Data.</b> A managerial presentation of Big Data concepts, landscape, and market trends. An introduction to Big Data marketing and business-oriented use cases, applications, and data sources (internal/external).	1,2	

2	<b>Introduction to Big Data architecture.</b> The enterprise perspective to Big Data initiatives. The new paradigm of an EDW (enterprise data warehouse).	1,2	
3-4	<b>Distributed Processing - Map Reduce.</b> Introduction to the Map Reduce paradigm. Demonstration of architectural concepts including presentation of basic examples (e.g., words count), design considerations, more advanced implementations: paralleling a data mining algorithm and implementing a join query.	3	
5-6	<b>Distributed Storage - DFS, Hadoop and SPARK.</b> Overview of the main design principles for managing and storing massive data sets at scale (demonstrated by HDFS). Presenting the underlying Hadoop architecture, technology stack including the Hadoop run modes and job types.	4,5	
7	<b>Spark I</b> – basic PySpark, structured and semi-structured data handling.	8,9	TA (Hands-on)
8	<b>Spark II</b> – SparkSQL, Advanced Spark	10,11	TA (Hands-on)
9-10	<b>Storage.</b> Presenting new concepts of data management including data lake, scalable relational databases. Introduction to NoSQL databases, Column-Based DB, Document store, key-value store and Graph DBs.	8,9,10,11	
11	<b>Pub/Sub systems and Real time data</b> – Introduction of Real time analysis, event stream concept. Kafka vs RabbitMQ vs Spark.	6,7	
12	<b>The future of big data</b> – Regulation – GDPR, ethics and selected UC	12	

\*התכנית הינה בסיס לשינויים.

בכל שבוע המצגת תכיל הפניה לקריאה הרלבנטית

## קריאת חובה

1. Lam, Chuck. Hadoop in action. Manning Publications Co., 2010.
2. H. Garcia-Molina, J. D. Ullman, J. Widom, Database Systems: The Complete Book. Second Edition. Pearson Prentice Hall, 2009.
3. D. Ullman and A. Rajaraman. Mining Massive Datasets. Cambridge University Press, UK 2012.
4. Provost, F., and Fawcett, T. Data Science for Business. O'Reilly Media, USA 2013.

1. Lynch, Clifford, (2008). Big data: How do your data grow?" *Nature*, 455(7209), pp. 28-29.
2. LaValle, Steve and Lesser, Eric and Shockley, Rebecca and Hopkins, Michael S and Kruschwitz, Nina. (2013). "Big data, analytics and the path from insights to value." *MIT Sloan Management Review*, 21.
3. Yang, Hung-chih and Dasdan, Ali and Hsiao, Ruey-Lung and Parker, D Stott. (2007). "Map-reduce-merge: simplified relational data processing on large clusters." *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 1029- 1040.
4. Shvachko, Konstantin and Kuang, Hairong and Radia, Sanjay and Chansler, Robert (2010). "The hadoop distributed file system." *Mass Storage Systems and Technologies (MSST)*.
5. Borthakur, D. (2007). "The hadoop distributed file system: Architecture and design." *Hadoop Project Website*, volume 21.
6. Thasos, Ashish and Sarma, Joydeep Sen and Jain, Namit and Shao, Zheng and Chakka, Prasad and Anthony, Suresh and Liu, Hao and Wyckoff, Pete and Murthy, Raghotham. (2009). "Hive: a warehousing solution over a map-reduce framework." *Proceedings of the VLDB Endowment*, 2(2), pp. 1626-1629.
7. Khan, Nawsher and Yaqoob, Ibrar and Hashem, Ibrahim Abaker Targio and Inayat, Zakira and Ali, Waleed Kamaleldin Mahmoud and Alam, Muhammad and Shiraz, Muhammad and Gani, Abdullah (2011). Big Data: Survey, Technologies, Opportunities, and Challenges.
8. Brewer, Eric (2012). "Pushing the CAP: Strategies for consistency and availability," *Computer*, 45(2), pp. 23- 29.
9. Cattell, Rick (2011). "Scalable SQL and NoSQL data stores" *ACM SIGMOD*, 39(4) pp. 12-27;
10. Han, Jing and Haihong, E and Le, Guan and Du, Jian. (2011). Survey on NoSQL database", *Pervasive computing and applications (ICPCA)*.
11. Stonebraker, Michael. 9201). SQL databases v. NoSQL databases." *Communications of the ACM*, 53(4), pp. 10-11.
12. Borner, K. and Poley, D.E. (2014). *Visual Insights -A Practical Guide to Making Sense of Data*. The MIT Press.