

## תואר שני

### 1242.3270.01 – מבוא לטכנולוגיות נתוני עתק Introduction to Big Data Technologies

**(דרישות במקביל):** מדע הנתונים למנהל עסקים או מבוא לישומי דאטה במנהל עסקים או נושאים מתקדמים בכריית מידע וגילוי ידע או גילוי ידע ורשתות נוירונים או נושאים מתקדמים במדע הנתונים למנהל עסקים)

#### סמסטר א – תשפ"א - מחצית שנייה

קבוצה	יום בשבוע	שעה	כיתה	תאריך בחינה	מרצה	דואר אלקטרוני	טלפון
03	יום ה'	21:30-18:45			מר משה קפלן	moshe.kaplan+mba@brightaqua.com	

שעת קבלה – בתיאום מראש

תאריכי מפגשים:

6.12 עד סוף הסמסטר

**\*הערה:** בנוסף, לסטודנטים ללא רקע בסיסי בתכנות וSQL מומלץ ללמוד קודם או במקביל את הקורס "טיפול יישומי בנתוני אנליטיקה עסקית"

#### היקף הלימודים

1 י"ס

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

#### תיאור הקורס

בעולם העסקי של היום נאספות ונאגרות כמויות מידע הגדלות בקצב מסחרר. היכולת לשלוט ביעילות בכמות (Volume) שטף (Velocity), ומגוון (Variety) סוגי הנתונים, הופכת להיות משאב ארגוני בעל ערך בסביבה עסקית גלובלית ותחרותית. יכולת זו מאפשרת לחברות להשיג תועלות משמעותיות בתחומים כגון שיווק, מכירות, ניהול מלאי, ועוד. על מנת להפיק תובנות אנליטיות בעלות ערך לארגון, נדרש לבצע פעולות קריטיות הכוללות איסוף, עיבוד, איחסון וניתוח נתונים במימדי ענק. לשם כך נדרשת תמיכה במגוון של טכנולוגיות חדשניות הקרויות טכנולוגיות נתוני עתק (Big Data technologies). קורס זה סוקר טכנולוגיות המאפשרות מתן מענה טכנולוגי הולם לדרישות Big Data ואשר אינם נתמכים בעזרת כלים סטנדרטיים בעולם טכנולוגיות המידע הארגוניים. במהלך הקורס ייסקרו שיטות לעיבוד מקבילי Map Reduce (, יכולת אחסון ועיבוד מבוצרות ומקביליות (Hadoop, DFS), הרחבות המאפשרות גישה נוחה לנתוני עתק (Pig, Hive etc). כמו כן ייבחנו טכנולוגיות חדשניות של מסדי נתונים (NoSQL, In-Memory Databases)

המותאמים לסביבה המאופיינת בכמות ובמגוון סוגי נתונים (structured/unstructured), תוך התמקדות בתכונותיהם המאפשרות התאמה לניתוח נתוני עתק. הקורס יתמקד בהצגת פתרונות ישומיים לתחום טכנולוגיות נתוני עתק (Big Data) בסביבה עסקית ושיווקית. בתום הקורס משתתפיו יוכלו לעמוד על הפוטנציאל הטמון במגוון הפלטפורמות והטכנולוגיות של טכנולוגיות המידע לצורך פתרון בעיות אנליטיקה עסקית מבוססי נתוני עתק. הקורס יספק בסיס תיאורטי שיאפשר העמקה בתחום וכן התנסות מעשית בשימוש ראשוני בטכנולוגיות נתוני עתק בענן (Cloud).

## תפוקות למידה

הקורס מורכב מהרצאות בכיתה שיועברו ע"י המרצה, ומהרצאות מרצה אורח. כמו כן, יכלול הקורס פרוייקט בו יישמו התלמידים את הנלמד בקורס באמצעות שימוש בטכנולוגיית ניתוח לנתוני עתק.

## הערכת הסטודנט בקורס והרכב הציון

אחוז	מטלה	תאריך	גודל קבוצה/ הערות
25%	פרוייקט מסכם		הפרוייקט המסכם יכלול פיתוח קוד ב - Python (PySpark מעל Spark) ו/או הרצאה על טכנולוגיה בתחום נתוני עתק
75%	בחינה		

\* עקב מאפייני תקופת הקורס יתכן כי תמהיל הערכת הסטודנט יהיה שונה

\* עפ"י תקנון האוניברסיטה תלמיד חייב להיות נוכח בכל השיעורים (סעיף 5).

\* מועד הבחינה יפורסם באתר הפקולטה- לוח בחינות.

\* 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)

## פירוט המטלות בקורס

מטלת הקורס (פרוייקט מסכם) תחייב לימוד עצמי. במסגרת הפרוייקט יחולקו הסטודנטים לקבוצות לימוד. ציון מעבר בקורס מותנה בציון מעבר בבחינה.

קיימת אפשרות שחלק מהרצאות הקורס יוחלפו בהרצאות אורח /הרצאות על נושאים אקטואליים בהתאם לשיקול דעת המרצה וראש התוכנית.

הרצאת אורח מחייבת נוכחות חובה.

הקורס ברובו יעסוק בהיבטים תיאורטיים, יחד עם זאת במהלך הפרוייקט הסטודנטים ידרשו לממש מספר רוטינות בקוד תוכנה בסיסי - עבורן תתקבל הדרכה במהלך הקורס.

## מדיניות שמירה על טווח ציונים

החל משנה"ל תשס"ט מונהגת בפקולטה מדיניות שמירה על טווח ציונים בקורסי התואר השני. עקרונות השיטה חלים על כל קורסי התואר השני, ומדיניות השמירה על טווח הציונים תיושם לגבי הציון הסופי בקורס זה.

מידע נוסף בנושא זה מתפרסם בהרחבה באתר הפקולטה.

<http://recanati.tau.ac.il/masters/yedion/2014-15/mba-rules-tests>

## הערכת הקורס ע"י הסטודנטים

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

## תכנית הקורס \*

Topic	Topic	Optional Reading
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	<b>Storage I</b> – Presenting new concepts of data management including: Data Lake, Scalable relational databases, CAP theorem, ACID vs, BASE, In-memory databases and transaction processing.	8,9
4	<b>Storage II</b> - An introduction to NoSQL databases, Column DB, Document store, Key-value store and Graph databases.	10,11
5	<b>Big Data Visualization Techniques</b> – “WHEN” (temporal data), “WHERE” (geospatial data), “WHAT” (topical data) and “WITH WHOM” (tree and network data). Introduction to methodologies and tools (e.g., Gephi or Elastic stack).	12
6	<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
7	<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
8	<b>Data Management and the Hadoop ecosystem</b> – Introduction of Pig for expressing data analysis and infrastructure processes; Hive for viewing data in HDFS; and HCatalog as the main concept of metadata representation of the Hadoop environment. <b>Optionally:</b> same topics in the Spark ecosystem.	6,7
9	<b>Big Data on The Cloud</b> – Walkthrough tutorial. TBD	

התכנית הינה בסיס לשינויים.

?

## 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.

## קריאת רשות

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.