

בית הספר למוסמכים במינהל עסקים ע"ש ליאון רקנאטי

## תואר שני

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

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

#### מסטר א' – תשפ"ב - מחצית שנייה

קבוצה	יום בשבוע	שעה	כיתה	תאריך בחינה	מרצה	דואר אלקטרוני	טלפון
01	יום ד'	21:30-18:45 מחצית שנייה	408	כמפורט בלוח הבחינות	ד"ר משה אונגר	mosheunger@tauex.tau.ac.il	03-6408112

שעת קבלה – בתיאום מראש.

תאריכי מפגשים: 28.11.21 – 16.1.22

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

#### היקף הלימודים

1 י"ס

ECTS = 4 י"ס – (European Credit Transfer and Accumulation System), 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 בעיילות בכמות (Volume), שטף (Variety) ומגוון (Velocity) 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 אשר אינם נתמכים בעזרת כלים סטנדרטיים בעולם טכנולוגיות המידע הארגוניים. במהלך הקורס ייסקרו שיטות לעיבוד מקבילי (Map Reduce), יכולת אחסון ועיבוד מבוזרות ומקביליות (Hadoop, DFS), הרחבות המאפשרות גישה נוחה לנתונים, כגון Pig ו-Hive. כמו כן, נדון בטכנולוגיות עדכניות של מסדי נתונים, כגון: NoSQL ו-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 בסביבה עסקית ושיווקית. בתום הקורס הסטודנטים 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 בטכנולוגיית נתוני עתק.

## 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

הקורס מורכב מהרצאות בכיתה שיועברו ע"י המרצה ומתרגולים שיהוו התנסות מעשית בשימוש בטכנולוגיית נתוני עתק. כמו כן, יכלול הקורס פרוייקט מסכם (זוגות) בו הסטודנטים יישמו פתרונות לבעיה עסקית באמצעות שימוש בטכנולוגיית ניתוח לנתוני עתק. הפרויקט המסכם יכלול פיתוח קוד ב-Python-Spark.

אחוז	מטלה	תאריך	גודל קבוצה/ הערות
25%	פרוייקט מסכם	הגשה עד שבוע מיום הלימודים האחרון בסמסטר	הגשה בזוגות, הפרוייקט המסכם יכלול פיתוח קוד ב-Python-Spark מעל PySpark
75%	בחינה	ייקבע בהתאם ללוח הבחינות	חובת מעבר בציון של 60 לפחות

### הערות לגבי ציונים:

- יתכן בונוס אישי של עד 5% מהציון על השתתפות ותרומה משמעותית לדיון בהרצאות.
- 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).
- על הסטודנט לקבל ציון 60 ומעלה במבחן המסכם על מנת להיות זכאי לקבלת ציון סופי ולהשלים את הקורס.

## פירוט המטלות בקורס

מטלת הקורס (פרוייקט מסכם) תחייב לימוד עצמי. במסגרת הפרוייקט יחולקו הסטודנטים לזוגות. הקורס ברובו יעסוק בהיבטים תיאורטיים, יחד עם זאת בפרוייקט המסכם הסטודנטים ידרשו לממש מספר רוטינות בקוד תוכנה בסיסי - עבורן תתקבל הדרכה במהלך הקורס.

חלק מהבחינה תכלול מימוש קוד תוכנה בסיסי (בדומה לפרוייקט המסכם).

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

## 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.

## 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 והתרגולים יופיעו באתר 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	<b>Storage I</b> – Presenting new concepts of data management including: DataLake, Scalable relational databases, CAP theorem, ACID vs, BASE, In- memory databases and transaction processing.	8,9	
4-5	<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	
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
8	<b>Storage II</b> - An introduction to NoSQL databases, Column DB, Documentstore, Key-value store and Graph databases.	10,11	
9	<b>Spark II</b> – SparkSQL, Advanced Spark	10,11	TA
10	<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	
11	<b>Pub/Sub systems and Real time data</b> – Introduction of	6,7	

	Real time analysis, event stream concept. Kafaka vs RabbitMQ vs Spark.		
12	<b>The future of big data</b> – Regulation – GDPR, ethics of warehousing and selected UC	<b>12</b>	

\*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. (2011). 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.