אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי הטבע - המחלקה למדעי המחשב
סמסטר ב' תשע"ד

סילבוס קורס

שם הקורס:	**תכנות מערכות מבוזרות**
שם הקורס באנגלית:	**Distributed System Programming: Scale Out with Cloud Computing and Map-Reduce**
מס' קורס:	**202-1-5391**
סוג קורס:	**קורס בחירה**
נק"ז:	**4.0**
מרצה הקורס:	**ד"ר מנחם אדלר**
דרישות קדם	**202-1-2011 אוטומטים ושפות פורמאליות** **202-1-2031 תכנות מערכות**

סילבוס באנגלית

Modern Web-scale applications (e.g., Facebook, Twitter, Google Docs) must face technical challenges that derive from their scale:

- Scalability: the possibility to grow as the user-base and data-size handled by the application grows to hundreds of millions of users and petabytes of data.
- High-Availability: the capacity to provide service to users even when part of the infrastructure (CPUs, Networks, Disks) become unaccessible in an intermittent or permanent manner.

The way to address these requirements is to develop loosely distributed applications that can operate in a "cloud-like" runtime environment. This course introduces basic theory behind such massively distributed applications and modern programming tools that constitute an emerging infrastructure for distributed applications, as well as algorithm design in this environment, demonstrated by several classic problems in the field of natural language processing.

מטרת ונושא הקורס

אפליקציות מודרניות - דוגמת Facebook, Twitter, Google Docs-, המבוססות על מאגר גדול ומבוזר של נתונים, חייבות להתמודד עם אתגרים טכניים שונים. ובפרט:

- הסתלמות: עמידה בגידול מתמיד של משתמשים (מאות מיליונים) ונתונים (peta-bytes) במערכת.
- זמינות גבוהה: יכולת מתן שירותים למשתמשים, גם כאשר חלק מהתשתית (מעבדים, רשת, כוננים) אינו זמין.

כדי להתמודד עם דרישות אלו, נדרש לעצב אפליקציה מבוזרת, הכוללת מספר רב של תהליכים, הרצים במקביל תחת 'ענן' של יחידות חישוב שונות, והממלאים יחדיו את המשימה העומדת בבסיס האפליקציה.
מטרת הקורס הינן לימוד התשתית התאורטית והמעשית של עיצוב מערכות מבוזרות לעיבוד אוסף נתונים גדול ומבוזר:

- עבודה מול ענן של יחידות חישוב ומאגרי נתונים מבוזרים
- עיצוב אלגוריתמים מבוזרים בתבנית map-reduce, ובפרט הכרות עם אלגוריתמים קלאסיים של אחזור מידע ועיבוד טקסט.
- מימוש מערכות מבוזרות לעיבוד טקסט בסביבת ההרצה Hadoop

נושאי ההרצאות

1. מבוא
- הצורך במערכות מבוזרות
- תכונות של מערכות מבוזרות
- ארכיטקטורת clouds
- מערכת קבצים מבוזרת

2. סביבת ההרצה Hadoop
- הארכיטקטורה של סביבת ההרצה Hadoop
- תבנית map/reduce
- תכנות בתבנית map-reduce בסביבת ההרצה Hadoop
- הוספת שכבת caching לסביבה הקיימת

3. עיצוב אלגוריתמים מבוזרים
- טכניקות יסוד: Relational joins ,Order inversion, Secondary sorting,Local aggregation
- עיבוד טקסט
  - מודל שפה
  - תיוג חלקי דיבר
    - מודל מרקוב מבוזר
    - לימוד לא-מונחה של תנאי התחלה
  - ניתוח תחבירי
    - מבנה פסוקיות, מבנה תלויות
    - בעיית הניתוח התחבירי
    - לימוד מונחה של מבנה תלויות כבעיית סייוג, גרסה מקבילית של אלגוריתם perceptorn
  - סיווג טקסטים
    - עיצוב מבוזר בתבנית, map-reduce, של k-means
    - עיצוב מבוזר, בתבנית map-reduce, של LDA
- עיבוד מבוזר של גרף
  - מציאת המסלול הקצר לקדקוד נתון
  - קביעת page rank לדפים ברשת

4. תאוריית map-reduce
- אופטימיזציית פעולת join בסביבת map-reduce
- חסמים עליונים ותחתונים לחישוב map-reduce

דרישות הקורס

התנסות מודרכת ועצמאית בבניית תשתית ושימוש בשרותי cloud
שני תרגילי תכנות
פרויקט סופי

מרכיבי ציון הקורס

תרגילים: 50%
פרויקט סופי: 50%

ספרות הקורס

Chuck Lam, Hadoop in Action, Manning Publications, Nov 2010.

Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Morgan and Claypool Publishers, April 2010.

Moshe Levinger, Uzi Ornan, and Alon Itai. Learning morpholexical probabilities from an untagged corpus with an application to Hebrew. Computational
Linguistics, 2001, 21:383–404.

Yoav Goldberg, Meni Adler and Michael Elhadad, EM Can Find Pretty Good HMM POS-Taggers (When Given a Good Start), ACL 2008.

Ryan McDonald, Keith Hall and Gideon Mann, Distributed Training Strategies for the Structured Perceptron, NAACL, 2010

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, The Page-Rank citation ranking: Bringing order to the Web, Stanford Digital Library Working Paper SIDL-WP-1999-0120, Stanford University, 1999.

Amy N. Langville, Carl Dean Meyer, Google page rank and beyond, Princeton University Press, 2006.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce. ACM International Conference on World Wide Web, 2012.

F. N. Afrati and J. D. Ulman, Optimizing joins in a map-reduce environment, EDBT, March, 2010.

Foto N. Afrati, Anish Das Sarma, Semih Salihoglu and Jeffrey D. Ullman, Upper and Lower Bounds on the Cost of a Map-Reduce Computation (submitted, 2012).