הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

Coller School
of Management
Tel Aviv University

בית הספר למוסמכים במינהל עסקים ע"ש ליאון רקנאטי

פרויקט מסכם- סמסטר ב' תשפ"ה

מבוא לטכנולוגיות נתוני עתק

תאריך הגשה: 26.5.25

מטרת הפרויקט המסכם היא לתרגל באופן מעשי את הנלמד בקורס באמצעות מקרה בוחן עסקי עם נתונים עסקיים. בפרויקט המסכם תנתחו בקבוצות בעיות עסקיות בעולם מערכות מידע גיאוגרפיות (GIS) ותספקו יישום של בעיות אלו באמצעות טכנולוגיות נתוני עתק שנלמדו בקורס.

במידה ואתם מבצעים הנחות כלשהן על הנתונים הקיימים, נא לציין זאת בסעיף הרלבנטי.

ניקוד על סעיפי המימוש (מסומנים כ-***) יינתן על פי הנכונות, אופן היעילות (זמן ריצה) בו תתכננו את עיבוד הנתונים ונראות הקוד (שם משמעותי לנתונים, תיעוד וכו'). שימו לב כי ניתן להשתמש בכלי GenAI, אך ינתן ניקוד על יעילות הקוד בהתאם לנלמד בכיתה ובתרגולים ויצירתיות בפתרון.

יש להגיש באתר הקורס 2 קבצים:

- קובץ PDF שיכיל תשובות לשאלות פתוחות. שם הקובץ יורכב מת"ז של הסטודנטים המגישים באופן הבא: ID1_ID2.PDF.
- קובץ .ipynb שיכיל את הקוד לשאלות התכנותיות (ראו סעיפים המסומנים ב-***). יש לרשום בקובץ בהערה את מספר וסעיף השאלה שהקוד מתייחס אליה, כולל הנחות יסוד. שם הקובץ המוגש יהיה ID1_ID2.ipynb.

הנחיות נוספות למימוש סעיפי הקוד:

- שימו לב באילו סעיפים אתם נדרשים לממש קוד (מסומן ב-***) עם פונקציות מ-SparkSQL (מבנה מסוג pyspark.sql datafame) לעומת סעיפים עם פונקציות מ-PySpark (pyspark RDD using mapreduce).
- הפרידו בין הסעיפים, כך שכל תשובה תופיע בבלוק נפרד ב-Notebook. בראש כל בלוק כיתבו הערה שמציינת את מספר השאלה, מספר הסעיף וסוג המבנים למימוש (SparkSQL עם DF או PySpark עם RDD).
- אם התבקשתם לענות תשובה מספרית, הפעילו את פונקציית print על מנת להציג את התוצאה. אם מדובר בטבלה, השתמשו בפונקציית display.
- במידת הצורך הוסיפו בשאלות SparkSQL עמודות לחישובי ביניים לטבלאות.

נתוני עתק

נתוני הפרויקט מבוססים על מערכת המידע GIS של חברת Gowalla. הטבלה העיקרית הינה Gowalla check-ins והיא מייצגת ביקורים של משתמשים ב-PlaceIds, כלומר מקומות POIs (Points of Interest). הטבלאות הנוספות מהוות פרטים נוספים על המשתמשים, המקומות, ועוד ורצוי להשתמש בהן בסעיפים השונים בפרויקט.

שאלות פתוחות (8%)

1. (4%) ע"ב הקבצים השונים של החברה, מה סוג הנתונים שנשמרים ע"י החברה? האם הם structured, semi-structured או unstructured? נמקו היטב את תשובתכם.
2. (4%) על פי תוכן הנתונים והבנתכם את המטרות העסקיות של החברה, מה סוג הארכיטקטורה שאתם מציעים ל-Gowalla לשמור את נתוניה (Data Lake, Date Warehouse, Hybrid)? הסבירו היטב את תשובתכם.

3. (17%) ברצוננו לממש מספר ETL (או ELT – תלוי בחירה של הארכיטקטורה בשאלה 2) עם מבנים מסוג Dataframe ב-SparkSQL בלבד. תהליכי העיבוד יתבצעו באופן הבא (כל סעיף הוא ETL או ELT נפרד):

   א. (3%) ***יש לחלק את העמודה הבודדת datetime ל-5 עמודות נוספות בפורמט של יום (Day), חודש (Month), שנה (Year), weekday (כן/לא), dayOfTheWeek (1-7). הניחו כי שבת וראשון **הינם** ימי weekend. יש לוודא תקינות הערכים של העמודות החדשות.

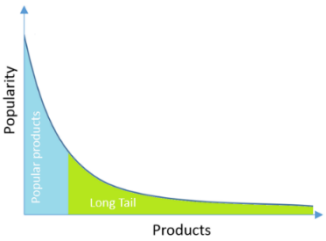
ב. (3%) *** יש להוסיף עמודה חדשה בשם last_place_visited אשר תחשב לכל שורת נתונים של משתמש את תאריך הביקור הקודם של המשתמש. תוכן העמודה יהיה ערך ה-datetime האחרון של אותו משתמש. במידה ואין לאותו משתמש ביקור קודם, יש להזין ערך None. לדוג׳, אם יש משמש שביקר בשלושה מקומות שונים בתאריכים 2010-06-28, 2010-06-24, 2010-05-01 אז ערכי העמודה יהיו 2010-06-24, 2010-05-01, ו- None בהתאמה.

ג. (8%) ***יש להוסיף לפחות 2 עמודות חדשות בשם last_x_place_visited (שימו לב ש-x מוזן על ידכם) אשר תחשב לכל שורת נתונים של משתמש את ה-x מאז הביקור הקודם של המשתמש. בחרו Xים מתאימים, כגון: מרחק שעבר מהביקור הקודם, האם המשתמש שינה קטגוריה או עיר מאז הביקור הקודם וכו'. מטרת סעיף זה לייצג את השינוי מאז הביקור הקודם כדי לתמוך בקבלת החלטות של חלוקת המשתמשים לקבוצות על בסיס דפוסי הביקורים שלהם באופן דינמי במערכת. רשמו בהערה בראש הבלוק מהן העמודות שבחרתם לבצע ונמקו בהתאם.

ד. (3%) ***יש לבצע פעולת "ניקוי" כלשהי של הנתונים כדי לתת תקפות (veracity) לנתונים. רשמו בהערה בראש הבלוק מה הפעולה שבחרתם לבצע ועל איזה מאגר נתונים ומהי ההנחה שביצעתם. פעולה לדוגמא יכולה להיות הסרת רשומות כפולות של אותו המשתמש על אותו מקום בתאריך זהה.

בשאלות הבאות (החל משאלה 4) השתמשו בנתונים המעודכנים לאחר מימוש תהליכי העיבוד משאלה 3.

4. (8%) ***ממשו את השאלה העסקית הבאה עם מבנים מסוג RDD ב-PySpark. עליכם לממש את הפונקציונאליות של פופולריות מקומות במערכת לטובת מערכת המלצה מבוססת פופולריות. בהינתן קטגוריה (gencat), שנה ו-k, יש להחזיר את top-k (באתר, למשל, k=5) המקומות הפופולריים, יחד עם מדד הפופולריות שלהם שמהווה את סך כמות הביקורים של משתמשים שונים באותה השנה.



5. (20%) ***ממשו את השאלה העסקית הבאה עם מבנים מסוג RDD ב-PySpark והוסיפו הסבר מילולי מתאים לתהליך שביצעתם: (**רמז**: בנו לכם עמודות מתאימות בסעיף 3ג'). עליכם לקבוע האם קיימת תופעה של long-tail בקטגוריית המסעדות (gencat=food) בכל אחת מ-2 הערים המרכזיות ב-dataset? נגדיר תופעת long-tail בעיר ושנה מסוימת, בה מעל 70% מהביקורים בשנה זו נמצאים ב- 20% מהמקומות הקיימים בעיר זו.

הצדיקו והסבירו את התוצאות שקיבלתם עם נתונים מתאימים לכל שנה. הסבירו בנוסף כיצד יצרתם מיפוי (התאמה) בין מסעדה לעיר שלה.

6. (47%) חברת Gowalla מעוניינת לענות על השאלה העסקית הבאה: כיצד בעלי עסקים (בעלי POIs) יכולים למצוא משתמשים בעלי ערך גבוה, בהתחשב בנתוני העבר שלהם במערכת, על מנת לספק תכנים שיווקיים (כגון קופונים) למשתמשים. לשם כך, חברת Gowalla מעוניינת ללמוד על מיקומי המשתמשים ולסווג משתמשים דומים על פי נתוני המיקום שלהם. החברה מעוניינת לפתח מערכת המלצה לבעלי העסקים שתהיה מבוססת על סיווג הדמיון הנ"ל, כך שבעת המלצה לבעל עסק POI, המערכת תוכל להשתמש בדמיון בין משתמשים על פי מאפייני המיקומים ונתונים נוספים הקיימים במערכת.

א. (6%) ***ממשו ETL (או ELT) אשר מחזיר ייצוג (מבנה חדש) עבור כל שורת נתונים של משתמש המהווה את כל המקומות שהוא ביקר בהם בעבר. עליכם להחליט כיצד לייצג באופן יעיל את נתוני המקומות של המשתמשים (טיפ: היעזרו בסעיף 3ג'). שימו לב ששיקולי אתיקה ופרטיות חשובים ביותר לחברה.

ב. (6%) הציעו אלגוריתם (רשמו בפסאודו-קוד) שמקבל Timestamp, userId ו-k ומחזיר את k המשתמשים הדומים ביותר למשתמש userId על פי נתוני העבר של המשתמשים במערכת שקודמים ל-timestmap. נמקו מהי לדעתכם מהירות החזרת התוצאות של האלגוריתם שלכם (real-time, near real-time, periodic, batch).

ג. (8%) הציעו אלגוריתם (רשמו בפסאודו-קוד) שמקבל Timestamp, placeId ו-k ומחזיר לכל בעל עסק (PlaceId) את k המשתמשים בעל הערך הגבוה ביותר עבורו, כך שהוא יוכל לשלוח אמצעי שיווק, כגון קופונים או מיילים. יש לסדר את המשתמשים לפי רמות החשיבות. נמקו מהי לדעתכם מהירות החזרת התוצאות של האלגוריתם שלכם (real-time, near real-time, periodic, batch).

ד. (7%) ברצוננו לתכנן Map-Reduce בטכנולוגיית Hadoop, כך שבאופן מקבילי ניתן יהיה לענות באותו התהליך מה סך כמות הביקורים של כל משתמש בכל עיר וקטגוריה של מקום וממוצע המרחקים בין מקומות שביקר בעיר וקטגוריה זו. הסבירו כמה שלבים נדרשים ב-Map-Reduce ומה יבצע כל Mapper ו-Reducer בכל שלב.

ה. (15%) ***ממשו את האלגוריתמים שהצעתם בסעיף ב' ו-ג' עם מבנים מסוג RDD ב-PySpark.

ו. (5%) הציעו דרך לבחון את יעילות האלגוריתם שלכם. איך תחליטו האם האלגוריתם מוצלח ונותן value לארגון? נמקו תשובתכם.

בהצלחה!