כריית מידע מנתונים גדולים **52019/52002 HUJI** - עבודת אמצע

תאריך להגשה: 2024-03-07

הוראות: יש להגיש את העבודה בזוגות או לבד.
משקל כל שאלה 10 נקודות. יש בסה"כ 11 שאלות מתוכן יש לענות על 10. עבור תלמידים שיענו על כל השאלות, תושמט התשובה עם הציון הנמוך ביותר בחישוב ציון העבודה.
את הפתרון לחלק הראשון הגישו בקובץ טקסט בשם: MidTerm_2023_24_SQL_<ID1>_<ID2>.txt.
את הקובץ יש להכין מתוך הקובץ MidTerm_52019_2023_24_SQL_Template.txt הנמצא במודל ולמלא בו רק בחלקים המיועדים לכך, כולל ID. אין לשנות שורות אחרות.
את הפתרון לחלק השני הגישו בקובץ טקסט בשם: MidTerm_2023_24_unix_<ID1>_<ID2>.txt.
את הקובץ יש להכין מתוך הקובץ MidTerm_2023_24_unix_Template.txt הנמצא במודל ולמלא בו רק בחלקים המיועדים לכך, כולל ID. אין לשנות שורות אחרות.
בנוסף בשאלות 4,5 בחלק השני יש לצרף לפתרון קבצים נוספים בהתאם להוראות.
את כל קבצי הפתרון יש לדחוס לקובץ zip בשם MidTerm_2023_24_<ID1>_<ID2>.zip ולהעלותו למודל.
בהצלחה

חלק ראשון: שאילתות SQL **במסד נתונים גדול** BigQuery
בחלק זה נשתמש ב-BigQuery של google cloud . הוראות לעבודה ב-BigQuery ניתנו בשיעור ובסרטון הדרכה במודל. אינפורמציה נוספת נמצאת כאן: BigQuery: Cloud Data Warehouse
מספר דברים שיש לשים לב אליהם:

- שירות BigQuery מאפשר גישה לנתונים בסקאלה גדולה של petabytes באמצעות פקודות SQL
- השימוש שיש לכם בגישה חינמית הוא מוגבל ב-TB בחודש. אנא שימו לב כשאתם מבצעים שאילתות לא לבצע שאילתות הפונות למספר רב של נתונים. בפרט, אין להשתמש בפקודה * SELECT אלא יש לשלוף את העמודות המתאימות. תוכלו לראות את גודל הנתונים שייסרק עבור כל שאילתה לאחר כתיבתה (אך לפני הרצתה) כאשר ייכתב לכם:
  "This query will process X MiB/GiB when run"
- בתרגיל זה נשתמש ב-database בשם stackoverflow. האתר Stack Overflow הוא אתר שימושי מאוד המהווה את קהילת האונליין הגדולה ביותר עבור מפתחי ומשתמשי קוד המעלים בו שאלות ותשובות לבעיות/אתגרים בהם הם נתקלים. ה-database מכיל פוסטים שהעלו משתמשי האתר וכן מידע על המשתמשים עצמם. בפרט, אנו נשתמש בשלוש הטבלאות הבאות:
  `bigquery-public-data.stackoverflow.posts_questions`
  `bigquery-public-data.stackoverflow.posts_answers`
  `bigquery-public-data.stackoverflow.users`
  כאשר הטבלה הראשונה מכילה רשומה עבר כל שאלה, השניה רשומה עבור כל תשובה והטבלה השלישית מכילה רשומה עבור כל משתמש.
- כאשר פונים לטבלאות ב-BigQuery יש להשתמש בגרשיים הפוכות בשם הטבלה ``
  (זה לא נחוץ ב-database SQL רגיל)
- עבור כל שאלה עליכם להדפיס כתשובה הן את שאילתת ה-SQL אותה ביצעתם, והן את התשובה המתקבלת אותה יש לצרף בפורמט json (יש ללחוץ על הכפתור JSON ולהעתיק את הטקסט המתקבל). אם נשאלתם שאלה נוספת (פרט לשאילתה והפלט שלה) יש לכתוב את התשובה בטקסט חופשי מתחת לפלט של השאילתה.

1. אנו מעוניינים למצוא את הפוסטים הטובים ביותר העוסקים בשפת ג'אוה-סקריפט. כתבו שאילתה השולפת את כל הפוסטים עבורם ה-tag מכיל את המילה "javascript" ללא חשיבות ל-case (כלומר גם כאשר כל המילה או חלקה היא באותיות אנגליות גדולות) יש להחזיר את ה-id,title,tag וכן מספר התשובות וה-score של כל פוסט כאשר הפוסטים ממויינים על פי ה-score בסדר יורד.
הערה: בשאלה זו טבלת הפלט המתקבלת היא ארוכה מאוד - כאן הדפיסו בנוסף לשאילתה רק את חמשת הפוסטים עם ה- score הגבוה ביותר המתקבלים כתשובה בפורמט json.
2. כתבו שאילתה המחזירה טבלה בת שתי שורות: בשורה הראשונה המספר הכולל של כל הפוסטים מהשאילתה הקודמת (כלומר עבורם ה-tag מכיל "javascript"), מספר הפוסטים שעבורם ניתנה לפחות תשובה אחת, מספר התשובות הממוצע, מספר הצפיות הממוצע וה-score הממוצע של כל הפוסטים, וכן התיאור "javascript". בשורה השניה אותם שדות (מספר כולל של פוסטים, פוסטים עם לפחות תשובה אחת, ...) אבל עבור הפוסטים העוסקים ב- java ולא ב- javascript , כלומר כך שה-tag מכיל את המחרוזת java ואינו מכיל את המחרוזת "javascript". התיאור יהיה "java". יש לתת שמות משמעותיים לששת השדות המוחזרים.
3. אנו רוצים לבצע סטטיסטיקה כדי לראות באיזה ימים בשבוע האתר עמוס יותר, כלומר להראות את מספר הפוסטים המועלים ביום ראשון, שני, .., שבת. חזרו על החלק של השאילתה מהסעיף הקודם המתאים לשורה של javascript (אין צורך לשאול על פוסטים ב-java ולא ב- javascript) , אבל הפעם כך שתוחזר התשובה עבור כל יום בשבוע בנפרד (כלומר שורה על כל יום בשבוע בטבלה המוחזרת). באיזה יום בשבוע מועלה מספר הפוסטים המינימלי/מקסימלי? האם יש הבדל באיכות ופופולריות הפוסטים בין הימים השונים? הדרכה: כדאי להשתמש ב-GROUP BY
4. כעת אנו רוצים למצוא את כל השאלות העוסקות בקשר בין javascript, לpython וכן את התשובות שלהן. כתבו שאילתה השולפת את כל השאלות עבורן ה-title מכיל את המילה " javascript " וכן את המילה "python" מטבלת השאלות , וכן את כל התשובות המתאימות לכל שאלה כזו מטבלת התשובות, כאשר כל שורה בטבלה המוחזרת תייצג זוג של (שאלה + תשובה). אם יש לשאלה מספר תשובות, אותה שאלה תופיע במספר שורות בטבלה המוחזרת. יש להחזיר את ה- id,title,tag ואת גוף הטקסט (ה-body) של השאלה והתשובה. עבור ה-Body יש להסיר את כל תווי ירידת השורה '\n' אחרת זה ייצור בעיות בהמשך העבודה עם הקובץ.
ציינו את מספר השורות המתקבל ושימרו את הטבלה הזו בקובץ בשם stackoverflow_javascript_python_qa.csv . הציגו את 5 השורות הראשונות ממויינות על פי ה-ID של השאלות.
(**שימו לב:** זוהי שאילתה כבדה כי היא מחזירה את ה-body של שתי הטבלאות. הריצו את השאילתה קודם בלי שדות אלו ורק כאשר אתם בטוחים שהפקודה נכונה הוסיפו את שני ה-body לשאילתה).
5. כעת נניח שתאגיד אמריקאי רוצה לשכור עובדים בכל רחבי ארה"ב. כתבו שאילתה המחזירה את המשתמש/ת המבטיח/ה ביותר (על פי reputation) עבור כל מיקום (עיר) בארה"ב מבין המשתמשים בעיר זו שיש להם מומחיות ב-javascript. השמיטו ערים בהן אין אף משתמש/ת כזו. יש להחזיר את השדות about_me, website_url, reputation, display_name, location המכילים אינפורמציה רלוונטית על המשתמש/ת. מיינו לפי reputation בסדר יורד והציגו את 5 התוצאות הראשונות.
הדרכה: שימו לב שהמיקום ושפות התכנות ב-database נכתבו בטקסט חופשי ויכולים שלא להתאים במדוייק למחרוזת ספציפית - עליכם לדאוג לכך שלא תפספסו עובדים בשל כך. מומלץ להשתמש בפקודת CASE של SQL. ניתן להניח שהמיקום מכיל את ארה"ב (אך ייתכן שבצורות שונות), ושיש התייחסות בתיאור המשתמש ל-javascript.
6. לבסוף נרצה לבחון אם יש קשר בין מספר השאלות/תשובות של משתמש ל-reputation שלו. חלקו את המשתמשים לקבוצות על פי מספר השאלות בכפולות של 1000 (כלומר בין 0 ל999, בין 1000 ל-1999 וכו) ועבור כל קבוצה כזו חשבו את ה-reputation הממוצע. האם יש קשר בין מספר השאלות ל reputation הממוצע? בצעו שאילתה דומה עבור מספר התשובות (במקום שאלות). מה מסקנתכם?

חלק שני: פקודות unix **לעבודה עם קבצים גדולים**

עבור כל שאלה בחלק זה יש לכתוב בתשובה את פקודת ה-**unix /** מספר פקודות ה-**Unix** המבצעות את הנדרש בשאלה.

בנוסף, יש להריץ את הפקודה ולהעתיק גם את הפלט המתקבל לתשובתכם. בחלק מהשאלות הפלט הוא גדול – במקרה זה השתמשו בפקודת 'less' או 'head' והציגו את התחלת הפלט.

1. הורידו למחשב שלכם את הנתונים של stackoverflow מה-BigQuery משאלה מספר 4 לקובץ בשם stackoverflow_javascript_pyhon_qa.csv בפורמט csv. הדפיסו את 10 השורות הראשונות של הקובץ. השתמשו בפונקצית wc וציינו את מספר השורות, המילים והתווים בקובץ.
2. ספרו את מספר **השורות** השונות בהן מופיעה המילה pandas או המילה numpy בקובץ.
כעת ספרו את מספר **השאלות** השונות בהן מופיעה המילה pandas או המילה numpy בגוף השאלה בקובץ זה. שימו לב שאותה שאלה יכולה להופיע בכמה שורות.
3. חלקו בעזרת פקודות unix את הקובץ לקבצים שונים על פי שנת היצירה של השאלות (כלומר קובץ אחד לשאלות שנוצרו ב-2022, אחד ל-2021 ....), כללו את השנה בשמות הקבצים המתקבלים. כמה קבצים קיבלתם? הציגו את השורה הראשונה בכל קובץ .
4. כתבו תוכנית פייתון בשם count_top_diff_freq_words.py המקבלת כקלט **זוג** קבצי טקסט, מחשבת עבור כל אחד משניהם את השכיחות היחסית של כל המילים (מופרדות ע"י רווח), ומחזירה את k המילים עם **ההפרש** המקסימלי בשכיחות היחסית. יש להתעלם מה-case של הטקסט.
כלומר לדוגמה אם מריצים:
>python3 count_top_diff_freq_words.py file1.txt file2.txt 5
התוכנית תחזיר את 5 המילים שהשכיחות היחסית שלהן בfile1 הכי גבוהה ביחס לfile2.
(ניתן להיעזר בקוד של התוכנית count_words מתרגיל הרשות)
כעת בצעו בunix לולאה על הקבצים של השנים הנפרדות מהשאלה הקודמת והריצו את התוכנית על כל קובץ של שנה נפרדת, ביחס לקובץ הגדול המכיל את כל השנים. הציגו את 5 המילים הנפוצות ביותר בכל שנה ביחס לכלל השנים.
5. כתבו shell סקריפט בשם duplicate_and_zip.sh המבצע את הפעולות הבאות:

- יוצר קובץ חדש שהוא שכפול של הקובץ stackoverflow_javascript_pyhon_qa.csv 10 פעמים.
- ממיין את הקובץ המשוכפל.
- דוחס בנפרד את 3 הקבצים: הקובץ המקורי, המשוכפל, והמשוכפל לאחר מיון.
- מדפיס את גדלי הקבצים הדחוסים (עם ובלי השכפול) וכן את זמן הריצה של כל דחיסה בנפרד.

כעת, הריצו את הסקריפט באמצעות הקלאסטר hurcs. צרפו את פקודת ההרצה באמצעות sbatch, את ה-ID שה-job שלכם קיבל ואת קובץ הlog של הקלאסטר (בסיומת out.)
האם זמן הדחיסה של הקובץ המשוכפל היה גדול יותר מאשר פי 10 מזמן הדחיסה של הקובץ המקורי?
האם גודל הקובץ המשוכפל דחוס היה גדול יותר מאשר פי 10 מגודל הקובץ הדחוס המקורי?
מה לגבי הקובץ המשוכפל לאחר מיון? (עבור 2 השאלות).

בהצלחה