מכללת הדסה, החוג למדעי המחשב

אנליזה של ביג דאטה / חננאל פרל

סמסטר ב', תשפ"ד

תרגיל מתגלגל מסכם / 1.06.2024 - *עדכון 24.06.2024 - 23:00*

תאריך הגשה:

הגשת סופית של כל החלקים: יום שני 1.07.2024 בשעה 23:59

הצגה חובה פיזית פרונטלית בכיתה: בשיעור האחרון יום רביעי 3.07.2024 בשעה 17:00

(כל מי שלא יכול להגיע פיזית צריך לקבל אישור כתוב ממני)

הגשת אמצע של לפחות חלק א: יום שלישי 25.06 בשעה 23:59

מטרות התרגיל:

תרגול מתגלגל מסכם

הוראות:

מתוך הסיליבוס:

התרגילים בחציו השני של הסמסטר יהיו תרגיל מתגלגל שלמעשה יורחב ויתפתח לכדי תרגיל מסכם.

בשיעור האחרון התרגיל המסכם יוצג בכיתה במשך כ-10 דקות לזוג (כל אחד מתוך הזוג יציג למשך 5 דקות). הצגה זו של התרגיל תשוקלל בתוך הציון הסופי של התרגילים.

כל זוג יצטרך להגיש מסמך קצר המתאר ומסביר את התרגיל המסכם, יש לציין ולתאר בצורה ברורה את האחריות והעבודה של כל אחד מהסטודנטים במהלך התרגיל.

דרישות:

חיפוש באינטרנט ובחירת מאגר נתונים גדול – עשרות מיליוני שורות, הרבה עמודות, ומאות רבות של מגה.

המאגר יתכן ויופיע כקבצי CSV, יתכן ויהיו כמה קבצי CSV ביחד שיגיעו לגודל הרצוי.

כל זוג (או בודד באישור) צריך לבחור מאגר שונה.

יש לקבל אישור על המאגר לפני תחילת עבודה עליו.

הגשת בקשה לעבודה על מאגר תתבצע כאן:

https://docs.google.com/forms/d/e/1FAIpQLSdIUtWnvO4GCv4Y1PtcDmEeE07ggePDnM10lU2JHyX6Yap69w/viewform?usp=sf_link

יש לעקוב בלינק (כל תלמיד רואה את כל הבקשות בכיתה) הזה האם המאגר אושר או לא ולענות לשאלות לפי הצורך (בגיליון עצמו אין אפשרות עריכה יש רק אפשרות תגובה):

https://docs.google.com/spreadsheets/d/1nviWO8cwRnDq_u7bgB5m4Nd8V49NPpvJ8By3jNfrbxc/edit?usp=sharing

המאגר הזה ישמש אותך בתרגיל והוא יהיה כמעין ביג דאטה Warehouse

יש להגיש תאור של המאגר, מבחינה טכנית מה גודלו, כמה שורות? כמה עמודות? וכמה סך נפח המידע?
וגם לתאר את התוכן של המאגר מה יש במאגר? מה התוכן הכללי שלו..? מאיפה הוא הגיע וכו..
יש להוסיף תאור מערכת סכמטית תאורטית שאוספת מעכלת ומעבדת המידע לכדי צורתו הסופית (= תיאור מילולי של capture ingest store compute שיוביל למידע שהורדת)

שלב הבא יהיה להבין את הדאטה ולבחון אותו, לראות מה יש בו ומה אין בו. מה המאפיינים של הדאטה ומה אני יכול ללמוד מהם, אלו תובנות אני רואה.
נשאל שאלות על הדאטה (4-6 שאלות)
נתכוב שאילתות SQL שעונות על השאלות
נחשוב מה הסיפור שנרצה לספר? איך נספר אותו?
נציג ויזואליזציות כחלק מהסיפור..

מבחינת המערכת המרכיבים שלה יהיו ככה:

על הביג דאטה נעבוד עם SQL ע"י כל סט היכולות, פונקציות אגרגציה OLAP וכו
אפשר יהיה לעבוד עם DuckDB התקנה והרצה לוקלית, לינק: https://duckdb.org
או שאפשר יהיה לעבוד עם Big Query שהוא כלי בענן, לינק: https://cloud.google.com/bigquery?hl=en
תנסו להבין מה יהיה יותר נח לכם.
[שיקולים לדוגמא: לוקלית הכלי זמין תמיד אך מוגבל ביכולות של המחשב שלכם, בענן אם נעבוד בחינם יש מגבלה של כמות הרצות שאפשר לעשות, אך הכח מחשוב כמובן חזק הרבה יותר]
שימו לב השפת SQL שבכלים יותר עשירה מאשר SQLITE. ישנם תוספת של פונקציות וכדומה, כדאי לקרוא קצת ולהשתמש ביכולות אלו.

לאחר שהרצתם את כל השאילתות על הביג דאטה, את התוצאות "הקטנות" נשמור בתוך SQLITE המוכר, לינק: https://www.sqlite.org
השאילתות יחזירו דאטה קטן במגוון שיטות: אגרגציה, פילטור, Samples וכו' לפי מה שלמדנו. יש לבחור כמה מהשיטות ולא רק שיטה אחת.
הדאטה בייס הקטן יכול להכיל כמה טבלאות.
ניתן להשתמש ב Pandas או כל כלי אחר להעברת הדאטה בין החלקים (פנדס ישמש להעברה בין המקוקות מידע בתור "צינור" לא להריץ שם מניפולציות או משהו אחר)

נייצר לבסוף דאשבורד על ידי StreamLit לינק: https://streamlit.io

לכתוב שם את השאלות את הסיפור ואת התובנות

להראות כ-50 שורות מאפיינות לדוגמא מתוך כל הטבלאות המקוריות, עם צבעים. להשתמש ב st.dataframe ואת הצבעים על ידי df.style.

וגם להוסיף 4 גרפים:

הגרפים 1-4 יהיו כולם או חלקם של Matplotlib - https://matplotlib.org/

ואפשר 1-2 מתוך הארבע שיהיו של Seaborn - https://seaborn.pydata.org/

מציין שיש לבחור גרפים מסוג דו מימד 2D (לא תלת מימד)

חלק א:

1. לינק למקור הנתונים
2. נתונים טכניים על המאגר, כמה שורות כמה עמודות וכמה סך הנפח המידע
3. תואר המאגר באופן כללי, הסבר מה יש בו ומה אין בו, מאיפה הגיע, מה המאפיינים של הדאטה וכו
4. הסבר מפורט על עמודות חשובות
5. יש להוסיף תאור מערכת סכמטית תאורטית שאוספת מעכלת ומעבדת המידע לכדי צורתו הסופית (= תיאור מילולי של capture ingest store compute שיוביל למידע שהורדת)
6. תאור מילולי וגם הפקודות שהרצת כדי להעלות את הדאטה הגדול לתוך הביג דאטה Warehouse שבחרת
7. השאלות ששאלתא על הדאטה (4-6 שאלות) ומה אני יכול ללמוד מהם, אלו תובנות אני רואה.
8. את השאילתות SQL שעונות על השאלות. וגם השאילתה שבוחרת שורות לדוגמא מכל טבלה.
9. קוד פייתון שבפועל מריץ את השאילתות מהדאטהבייס הגדול ומכניס לקטן.
10. את הקובץ דאטהבייס הקטן יש להגיש.
11. לכתוב מה הסיפור שנרצה לספר? איך נספר אותו?
12. הסברים אלו שיטות "להקטנת" הדאטה נבחרו כל פעם ולמה.
13. הסברים אלו וויזואליזציות נבחרו, למה ומה רואים בהן..
14. את הקוד פייתון של האפליקצייה (אפשרי כמה קבצים) שמציג הדאשבורד עם הטבלאות לדוגמא ואת הויזואליזציות (לא לשכוח מקרא). הדאשבורד יכלול גם הסברים על הסיפור מה רואים וכו.
15. כל קובץ נוסף שצריך כדי שנוכל להריץ את הדאשבורד לוקלי אצלנו.
16. קובץ requirements.txt המכיל את כל חבילת הפייתון הנצרכות להרצה.
17. הסברים איך להריץ.
18. צילומי מסך של הדפים העיקריים בדאשבורד.
19. קטע קצר המתאר ומסביר את התרגיל המסכם, מבחינת ארכיטקטורה וקוד, כולל ציור סכמתי של המערכת.
20. תיאור בצורה ברורה את האחריות והעבודה של כל אחד מהסטודנטים במהלך התרגיל.
21. קובץ **README** – שמות המגישים, ורשימת כל הקבצים שהוגשו (כל הקבצים שבתוך הZIP), והסבר קצר מה יש בכל קובץ.

כל ההסברים והתאורים והתמונות צריכים להופיע בקובץ וורד (docx) אחד.

ההגשה תהיה קובץ ZIP עם כל מה שמבוקש כאן ברשימה.

יש להגיש את הקובץ ZIP המלא כולו במודול.

אם הגעת למגבלה של 50 MB ואינך מצליח להגיש הכל במודול, אז:

א. נא להגיש את הקובץ ZIP המלא והגדול מעל 50 MB כאן:

https://docs.google.com/forms/d/e/1FAIpQLSfOnk3vPXkl8XI8sfHE3eddfWSNDJI3Aj94SKJ8RA5jblHqdg/viewform?usp=sf_link

ב. כמו כן, יש עדיין להגיש במודל את כל קבצי התרגיל למעט הדאטה בייס, מכווצים ב ZIP.

ג. אם הקובץ ZIP שלך קטן מ 50 MB – אין להגיש בלינק הזה, אלא להגיש הכל רק במודול!

הגשה חלק ב:

- תוספת של 2 גרפים נוספים, שיהיו אינטראקטיביים, ז"א היוזר יוכל לבחור משהו במסך ואז לקבל גרף מותאם אליו, כזכור התוכן של הגרפים, גם האינטרקטיביים, ילקח מהדאטה בייס הקטן SQLITE.
(אפשרי להפוך את אחד או שניים הגרפים מתוך ה-4 הקודמים לאינטרקטיבי, ולהוסיף חדשים, כך שלבסוף יהיו סה"כ 6 גרפים / ויזואליזציות)
אפשר שה-2 ויזואליזציות האלו יהיו מסוגים נוספים: טבלה עם צבעים, word cloud ענן מילים, מפות, ודברים נוספים לאו דווקא גרפים במובן בפשוט של המילה. ז"א אפשר דברים יותר יצירתיים לאו דווקא מתוך הסיפריה של Matplotlib. כל סיפריה שמוסיפים כאן, כמובן, יש לציין בקובץ requirements.txt.
- יש לוודא שמופיע בשאילתות גם קבוצות: GROUP BY | HAVING וגם חלון: OVER | PARTITION.
- יש לוודא שהדאטה הקטן אכן מכיל מינימום דאטה הנדרש להצגה בדאשבורד. רק שורות נצרכות, רק עמודות נצרכות ושאין כפילויות משמעותיות של דאטה בכמה טבלאות וכדומה..
- יש לייצר את הדאטה הקטן בצורה נקייה על ידי יצירת הטבלאות על דאטהבייס חדש (ללא מחיקות בדרך). אם כותבים ומוחקים הרבה פעמים הדאטהבייס יהיה הרבה יותר גדול ממה שצריך.
- יש להגיש רשימת כל הטבלאות שנמצאות בדאטה בייס הקטן (ה SQLITE) כמה שורת וכמה עמודות בכל טבלה, והסבר קצר מה יש בטבלה ולאיזה צורך.
- אם רואים שהביצועים של הדאשבורד על הדאטה הקטן מאד איטיים בטבלה או טבלאות, נא להוסיף אינדקסים ב SQLITE כדי לשפר הביצועים.
- אין להגיש את הקובץ (או קבצים) של הדאטה הגדול שהורדתם מהאינטרנט! (הגשה שלהם תוריד בציון..)
- יש להכין עבור כל קובץ (או קבצים) של הדאטה – קובץ מקביל עם אותו מספר עמודות אך רק עם 50 שורות ולהגיש אותו (או אותם).
- כל קובץ שמופיע בהגשה חייב להיות מתואר בקובץ README.

הגשה חלק ג:

- GRPAH DB:

אילו הייתם צריכים לשמור את הנתונים שהורדתם בחלק א בתור GRPAH DB. איך הם היו מסודרים שם.

יש לצייר סכמה (כמו מה שהפקודה הזו מציירת: CALL db.schema.visualization())

הסכמה המצוירת תכלול את ה NODES וגם את ה RELATIONSHIPS. יש צורך גם לציין מה יהיו ה PROPERTIES של כל אחד מהם.

יש צורך גם להסביר ולפרט באמצעות טקסט קצר את הסכמה ומה שנמצא בה.

אין חובה למדל את כל העמודות בקובץ, מספיק להראות כ-8 עמודות..

- FAKE DATA + SPARK:

יצירת קובץ נתונים פיקטיביים בצורת מסמך כמו JSON או משהו דומה (לא טבלה, לא CSV).

יש לייצר משהו רלוונטי לדאטה שנבחר בתחילת התרגיל.

הייצור יהיה ע"י קוד פייתון ושימוש בסיפריה Faker:

https://faker.readthedocs.io/en/master/

https://github.com/joke2k/faker/tree/master

Each of the generator properties (like name, address, and lorem) are called “fake”. A faker generator has many of them, packaged in “providers”.

Check the extended docs for a list of bundled providers and a list of community providers.

גודל הדאטה הנוצר צריך להיות בערך MB70.

אין להגיש את הקובץ דאטה המלא שיצרתם. יש לייצר גרסא מוקטנת של כ50 רשומות ולהגיש אותה.

את הקובץ המלא שייצרתם, יש לקרוא את הקובץ ולעבד באמצעות SPARK על גבי פייתון.

אפשר להתקין לוקלי במחשב שלכם או לעבוד על מחברת COLAB.

קוד להרצה בתוך COLAB:

```
!pip install pyspark
!pip install py4j

import pyspark
from pyspark.sql import DataFrame, SparkSession
import pyspark.sql.types as T
import pyspark.sql.functions as F

spark= SparkSession \
       .builder \
       .appName("Our First Spark Example") \
       .getOrCreate()
# to see Spark UI - https://stackoverflow.com/a/77312506
from google.colab import output
output.serve_kernel_port_as_window(4040, path='/jobs/index.html')
```

יש לשאול על הדאטה הסינטטי שיצרתם שאלה

לנתח ולעבד ע"י SPARK כדי לענות על השאלה בהתאם

לשמור בדאטה בייס "הקטן" SQLITE

ואז להציג בהתאם בדאשבורד, בדרך שתבחרו.

בנוסף יש להציג דוגמית של כמה שורות מתוך הדאטה הסינטטי שיצרתם.

יש צורך גם להגיש את הקוד פייתון של SPARK שהרצתם.

כמובן, גם כאן כל סיפריה שמשתמשים בה צריכה להיות ב requirements.txt.

כמובן, גם כאן כל קובץ שמוסיפים להגשה חייב להוסיף ב README עם הסבר קצר.

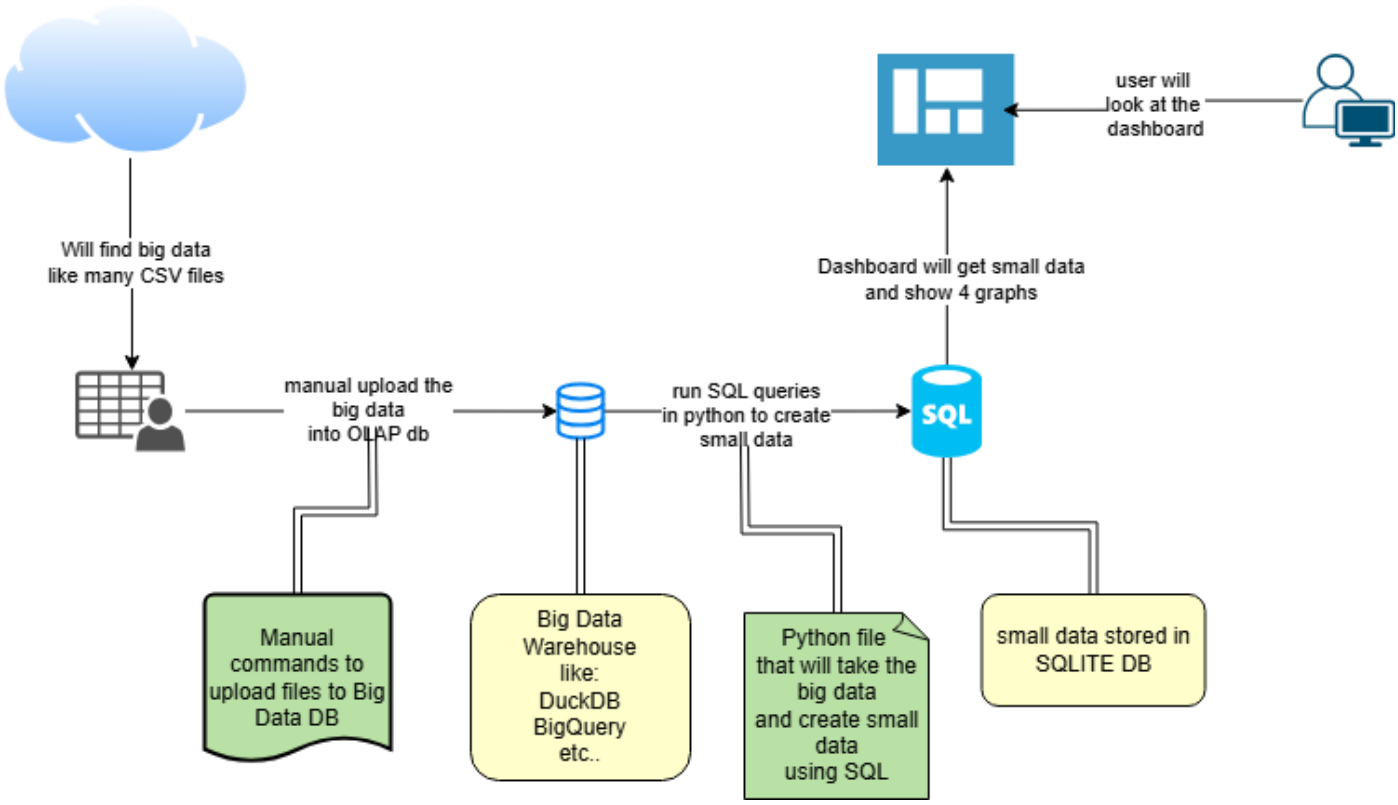
בסופו של דבר צריך להגיש את כל הקבצים מכל החלקים של התרגיל המתגלגל המסכם (ולא רק את הדלטאות ביחס להגשה הקודמת)

כמו בפרויקט בחיים האמיתיים, יתכנו שינויים ותיקונים במהלך התרגיל,

אנו עקבו אחר ההודעות!

סכמה:

סכמה של המערכת (ציור בעזרת https://www.drawio.com):



בהצלחה !!