|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **מס' קבוצה** | **שם הפרוייקט** | | | **תאריך הגשה** |
| ## | Website.com | | | DD/MM/YYYY |
| **מספרי תעודות הזהות של המגישים** | | | | |
| ######### | | ######### | ######### | |

**חלק ב' – הכנת תשתית הנתונים**

## **(יש להסיר את ההנחיות וההסברים לפני הגשת המטלה)**

## **חלק ב' בכללותו מהווה 10% מציון הקורס.**

## מטרת חלק ב' היא עיצוב והקמת תשתית הנתונים, על סמך האפיון הראשוני שבוצע בחלק א׳.

* בחלק זה אתם נדרשים לממש את מחסן הנתונים (Data Warehouse) בלבד.
* מימוש של מרכולי נתונים (Data Marts) וקוביות הOLAP ייעשה במידת הצורך בחלק הבא.

## **להלן מספר הנחיות כלליות, הנוגעות להגשת חלק זה:**

## את המטלה יש להגיש אלקטרונית במודל לתיבת ההגשה ע"י אחד מחברי הקבוצה. יש להגיש את המטלה במועד על פי ההנחיות שתפורסמנה באתר הקורס. הגשה מאוחרת מחייבת אישור מקדים מצוות הקורס.

## במתן הציון יושם דגש רב על מקוריות ופתרון בעיות יצירתי, וכן על הגשה מקצועית, נאה, ואפקטיבית. יש להשתמש בתבניות ההגשה שיפורסמו ב Moodle ולהקפיד על כתיבה בשפה רהוטה, מתומצתת, וברורה.

## אין לחרוג מהגבלת מספר העמודים בכל מרכיב במטלה. אי הקפדה על כך עלולה להוביל להורדה בציון.

* יש להגיש את כל המטלות הכתובות בדפוס בלבד - מטלות שיוגשו בכתב יד ו/או סרוקות, יוחזרו ללא בדיקה. יש להכין את כל התרשימים והשרטוטים בעזרת כלי תוכנה ובאנגלית. לא יתקבלו תרשימים סרוקים, ו/או כאלו שהוכנו בכתב יד
* חשוב להשתמש לאורך כל העבודה בשמות משמעותיים עבור שדות ועבור טבלאות. יש לצאת מנקודת הנחה כי התרשימים והטבלאות מיועדים לקורא שאינו מכיר ובקיא בסביבה העסקית.
* יש להגיש בחלק זה מספר קבצים:
  + קובץ Word זה, הכולל מענה על השאלות בההתאם לתבנית ההגשה.
    - עבור יישום אפיון תהליכי ETL (בסיסיים ומתקדמים), יש להוסיף צילומי מסך של תהליכי העבודה (flows). במידה והתהליך מורכב, יש לפרק אותו לשלבים ולהראות את השלבים בנפרד (בדומה למצגת בהדגמת ה-ETL).
    - במידה ונוספו שאילתות SQL או פרוצדורות שמורות, יש להציג אותן בשלב שהם נוספו בו.
    - במידה ונעשו שינויים של פילטר או שדות מחושבים, יש להציג צילום מסך שלהם.
    - שימו לב – על צילומי המסך להיות ברורים וקריאים.
  + קובץ Excel המתאר את תכנון ה-STT.
  + קבצי סרטונים להצגת היישומים המתקדמים, בהתאם להנחיות בסוף תבנית ההגשה

**ליישום חלק זה, נדרשות שתי הפלטפורמות הבאות:**

* שרת לבסיסי נתונים טבלאיים – SQL Server
* פלטפורמת אינטגרציית נתונים (Data Integration) – Tableau Prep

כלים אלו מותקנים במעבדות ההוראה וזמינים להורדה והתקנה במחשבים האישיים.

* הנחיות התקנה לכל הכלים הנדרשים ליישום חלקים ב' ו-ג' יפורסמו באתר הקורס (רישיון ומדריך התקנה לתוכנת ה- Tableau Prepנמצא תחת לשונית "מדריכים שימושיים" במודל. גרסת SQL-Server Express זמינה להורדה בחינם, ואת הוראות ההתקנה ניתן למצוא באתר הקורס.
* כללית, מומלץ ליישם את הפרויקט על כל חלקיו בעזרת המחשבים האישיים.

# (20%) יישום מקורות הנתונים ומחסן הנתונים הטבלאי

## **(כ- 1-2 עמודים)**

**עדכון סכמות הנתונים**

כמטלה מקדימה, יש לתקן את המודלים הטבלאיים שהוגשו בחלק א' - הן של מקורות הנתונים והן של מחסן הנתונים, על פי ההערות שניתנו על הגשת חלק א' של הפרויקט.

* בהגשה יש לכלול תרשים טבלאי עבור מקורות הנתונים (טבלאות מפורטות מתוקנות) וסכמת כוכב מתוקנת עבור מחסן הנתונים.
* לכל מודל נתונים – יש לתאר **בקצרה**, ברשימת נקודות תמציתית, מה השינויים העיקריים שנעשו בו לעומת הגרסה המקדימה שהוגשה בחלק א'.

**מימוש מקורות הנתונים – באמצעות שרת SQL Server**

יש ליצור את טבלאות בסיסי הנתונים המקוריים (מקור ראשי ומקור משני) בהתאם למודל המתוקן.

* כיון שהמקור הראשי משקף בסיס נתונים תפעולי – יש ליישם אותו בצורתו המנורמלת, עם כל אילוצי המפתחות הראשיים והזרים הנדרשים.
* בהגשה יש לכלול כנספח את קוד ה-SQL היוצר את הטבלאות (סדרת פעולות CREATE TABLE)

**אכלוס הנתונים**

בכל מקור נתונים, יש להזין לכל טבלה את הנתונים הנדרשים

* מספר רשומות גדול, שיאפשר ניתוח עשיר ומעניין בהמשך, אך לא מוגזם
  + סדר גודל של עשרות עד מאות רשומות בטבלאות המתארות ישויות
  + סדר גודל של אלפי עד עשרות-אלפי רשומות בטבלאות המתארות טרנזקציות עסקיות או קשרים.
* הנתונים צריכים להיות קריאים, מעניינים ובעלי הגיון עסקי.
  + אם הנתונים מופקים בסימולציה, אין להשתמש במחולל המייצר נתוני "ג'יבריש" חסרי משמעות.
  + מומלץ גם שלא להפיק משתני העובדה אקראיים לגמרי, אלא לתכנן הפקה שתוביל לאוסף נתונים בעל משמעות עסקית מעניינת (מגמות לאורך זמן, השפעה הדדית, שונות בין קבוצות, וכד')
  + ליצירת נתונים מעניינים ובעלי משמעות – מומלץ לתחקר את מגוון הפונקציות הקיימות ב-Excel
* בעת הכנת הנתונים והזנתם, חשוב לשמור על אילוצי המפתח הזר בין הטבלאות.
  + למשל; בטבלת הזמנות לא ניתן ליצור הזמנה עם לקוח, שלא קיים בטבלת לקוחות.

**מימוש מחסן הנתונים - באמצעות שרת SQL Server**

יש להקים גם את ארבעת בסיסי הנתונים הנדרשים למימוש תהליך ה-ETL ואת טבלאות מחסן הנתונים על פי המבנה המתוקן.

* לפני המימוש – **יש לקרוא היטב את ההנחיות** בסעיפים בהמשך.
  + יש לשים לב שההנחייה בחלק העוסק **היישומים הבסיסיים** היא לממש את מרבית טבלאות המימד כ"מימד משתנה מסוג 1", גם אם בתכנון המקורי הן עוצבו כ"טבלאות מימד משתנה מסוג 2".
  + בהמשך, בחלק העוסק **ביישומים המתקדמים**, אחת האפשרויות היא לממש את **אחד** המימדים כמימד משתנה מסוג 2.
  + בהתאם לבחירת המימדים שיעוצבו כמשתנים לאט מסוג 1 ואלו שיעוצבו כמשתנים מסוג 2, יש לעצב מחדש בהתאם הן את טבלאות המימד והן את המפתחות הזרים אליהן בטבלאות העובדה.
* בהגשה יש לכלול כנספח את קוד ה-SQL היוצר את טבלאות מחסן הנתונים (סדרת פעולות CREATE TABLE)

# (20%) אפיון תהליכי ה ETL

**(כ- 2-3 עמודים)**

**מיפוי מקורות ליעדים**

יש להציג תכנון STT (Source To Target) לפי הפורמט שהוצג בהרצאה והדוגמא באתר הקורס.

* התכנון יציג את מיפוי הטבלאות והשדות במחסן הנתונים למקורות הנתונים מהם יישלפו הנתונים, וכן את הטרנספורמציות הדרושות.
* אין צורך לכלול בתכנון זה את מימדי הזמן.
* על התכנון לשקף את כל התהליכים שימומשו בסעיפים הבאים – יישומי בסיס ויישומים מתקדמים.

**היערכות ואתגרים**

בנוסף להצגת תכנון ה-STT, יש לענות על השאלות הבאות:

* מהו הסדר הנכון של הרצת השלבים השונים בתהליך ה-ETL ?
  + יש לצרף תרשים של תהליך העל (Master), שמתאר את סדר הרצת השלבים של תהליך ה-ETL. המסמך אמור להיראות כמעין תרשים זרימה אשר מראה את סדר ההרצה של תהליכי העבודה שבניתם.
  + יש להוסיף הסבר קצר (בטבלה, או ברשימת נקודות) שמתאר את מטרת כל שלב ואת הדגשים החשובים ליישומו (פעולות ומשמעות).
  + אם קיימות תלויות בין שלבי התהליך – יש להסבירן בקצרה.
* במערכת האמת (כלומר, בהנחה שניאלץ בעתיד להתמודד עם כמויות נתונים בהיקף מלא) - אילו שלבי ETL צפויים לקחת זמן רב במיוחד, ומדוע ?
* אילו שלבים בתהליך עלולים להיכשל, וכיצד ניתן להתמודד עם כך ?

# (40%) מימוש תהליכי ה-ETL – יישומי הבסיס

**(כ- 2-3 עמודים)**

יש לפתח את תהליכי ה-ETL שישאבו וימירו את הנתונים משני המקורות ויטענו את הנתונים למחסן הנתונים, בהתאם לאפיון מודלי הנתונים ומיפוי ה STT בסעיפים הקודמים. במהלך מפגש ההגנה המסכם בסוף הסמסטר, תידרשו להציג את תהליכי ה-ETL שמימשתם.

* על תהליכי ה-ETL ליישם את כל השלבים הדרושים, על פי המתודולוגיה אשר נלמדה בכיתה - שאיבת הנתונים ממקורות הנתונים, המרתם למבנה הדרוש, ואכלוס הנתונים המומרים בטבלאות המתאימות.

**על יישומי הבסיס לכלול:**

* **טבלאות מימדי זמן:** מימדי זמן (תאריך ו/או שעה, בהתאם לצורך העסקי) חיונים בכל מחסן נתונים. את מימדי הזמן, על רשומות הנתונים הנדרשות בכל אחד מהם, יש לייצר ולאכלס ישירות במחסן הנתונים. באתר הקורס ניתן למצוא קוד ליצירת מימדי זמן – אך ניתן להשתמש גם בקוד חלופי, לבחירתכם.
* **טבלאות מימד:** בחלק זה, ככלל, יש לממש את המימדים כמשתנים לאט מסוג 1 (ללא שמירת גרסאות), גם אם בתכנון ה-DW חלק מהם עוצבו כמשתנים לאט מסוג 2.
  + יש לשים לב שבחלק הבא (יישומים מתקדמים) ניתנת אפשרות להמיר את אחת מטבלאות המימד, בטבלה שמשקפת עיצוב של המימד כמשתנה לאט מסוג 2.
  + במקרה של שימוש בסכמת פתית שלג – יש לתת את הדעת על סדר טעינה נכון.
* **טבלאות עובדה:** בחלק זה יש לממש את תהליכי הטעינה הן של טבלת העובדה הפרטנית והן של שתי טבלאות העובדה הסיכומיות.
  + עבור טבלת העובדה הפרטנית יש לממש טעינה אינקרמנטלית. יש להמחיש זאת במפגש ההגנה באמצעות דוגמא משכנעת – כיצד התהליך שולף אך ורק רשומות נתונים חדשות שנוספו למקור הנתונים, ומוסיף אותן בהתאם לרשומות הנתונים שכבר קיימות במחסן הנתונים.
  + עבור טבלאות העובדה הסיכומיות – אין צורך לממש טעינה אינקרמנטלית, אלא טעינה מלאה.
  + יש לתת את הדעת שחלק מהטבלאות מסתמכות על מקור נתונים אחד, וחלק על שני מקורות נתונים, ולתכנן בהתאם לכך את סדר הטעינה.

מימוש תהליכי ה-ETL יתבצע באמצעות ה- Tableau Prep , כלי מיזוג נתונים (Data Integration) המותקן במעבדות ההוראה וזמין להורדה והתקנה במחשבים האישיים בעזרת רישיון התקנה המפורסם במודל.

* במהלך הסמסטר, תינתן על הכלי הדרכה בסיסית. עם זאת, הכלי כולל אפשרויות מתקדמות רבות, והציפייה היא שהסטודנטים ישקיעו את המאמץ, וילמדו את הכלי לעומקו בכוחות עצמם
* הכלי יוצר סכמה גרפית של תהליכי ה-ETL על שלביו השונים. בהגשה יש לכלול צילומי מסך של הסכמות הגרפיות של מרכיבי התהליך (Flows) השונים, וכן יש חובה לסדר אותם על פי סדר ההרצה הנכון, סידור של תצלומים לא על פי סדר ההרצה הדרוש כפי שנלמד בקורס יגרור הורדת נקודות.

# (צבירה של עד 25%) מימוש תהליכי ה-ETL – יישומים מתקדמים

**(כ- 2-3 עמודים)**

הנכם נדרשים לממש שניים מתוך שלושת היישומים הבאים, הדורשים רמת מורכבות גבוהה יותר.

* לצורך ההתנסות, ניתן ואף מומלץ לממש את כל שלושת היישומים – אך במהלך מפגש ההגשה תהיה באפשרותכם להציג שניים מתוך השלושה בלבד, לבחירתכם.
* יש לשים לב שהיישום הראשון מבין השלושה מעניק ניקוד מעט יותר גבוה מהאחרים.
* לכל יישום שאתם מממשים יש לצרף סרטון הדגמה קצר, באורך של לא יותר מ-4 דקות לכל סרטון, בהתאם להנחיות. על הסרטון להיות תמציתיים. אין צורך להסביר את התהליכים אלא רק להראות שהם מתבצעים. במפגש ההגנה, נצפה יחד בסרטונים ותינתן לכם האפשרות להסביר אותם תוך כדי הצגתם.

**א. (15%) מימד משתנה לאט מסוג 2**

מעקב אחר שינויים לאורך זמן בנתוני טבלאות מימד עשוי להיות חיוני מבחינה עסקית ולניתוח נכון של הנתונים. עם זאת התחזוקה והעדכון של נתוני מימד משתנה לאט מסוג 2 מורכבת יותר, ודורשת הגדרה מתאימה של תהליכי ה-ETL שמזינים את טבלת המימד, וגם את אלו שמזינים את טבלאות העובדה המקושרות לאותה טבלת מימד.

ביישום זה, הדרישה היא להמיר את אחד ממשתני המימד שמומשו בחלק הקודם (לבחירתכם, למעט מימדי זמן/תאריך) למימד משתנה לאט מסוג 2, הכולל גם שמירת גרסאות.

* יש לעדכן בהתאם את סכמת מודל הנתונים של ה-DW - הן את סכמת טבלת המימד, ואת את סכמת טבלאות העובדה המקושרות למימד שנבחר.
* עבור המימד שנבחר - יש לתכנן ולממש באופן מלא תהליכי ETL, המטפלים כראוי בכל האתגרים הכרוכים במימוש מימד משתנה מסוג 2
  + הוספת גרסאות בעת זיהוי שינוי באחד השדות, ויצירת מפתח עזר בהתאם
  + קביעת קביעת ערכים מתאימים לשדות טווח התאריך (Valid From, Valid Until)
  + עדכון בהתאם של תהליכי הטעינה של טבלאות העובדה המקושרות למימד שנבחר – זיהוי הגרסה המתאימה, בהתחשב בהתאמת התאריך, ובהתאם לכך שימו בערך מפתח העזר הנדרש לקישור נכון של רשומת העובדה לגרסת המימד שמתאימה לה.

במהלך מפגש ההגנה המסכם תידרשו להמחיש באמצעות דוגמא משכנעת את פעולת תהליכי ה-ETL שמעדכנים את טבלת המימד ואת טבלאות העובדה – כיצד התהליך מזהה שינוי גרסה, ומעדכן בהתאם את הנתונים במחסן הנתונים.

**עבור יישום זה יש להראות בסרטון:**

* המחשה של הוספת רשומה חדשה עם תאריך תחילת תוקף של היום בו היא נוספה.
* עדכון רשומה: על הסרטון להראות שנוספה רשומה חדשה עם תחילת תוקף של היום בו היא נוספה, וסגירת הגרסה קודמת עם תאריך סוף תוקף של היום בו היא נוספה (או יום אחד לפני כן).
* התאמת גרסת המימד לרשומה חדשה שנוספת לטבלת העובדה הפרטנית: יש להראות שהרשומה שנוספה לטבלת העובדה אכן מפנה לגרסה המתאימה של הממד, בהתאם לטווח התאריכים.

**ב. (10%) טבלת עובדה מסוג "תמונת מצב" (Snapshot)**

טבלת עובדה מסוג "תמונת מצב" (Snapshot) נדרשת במצבים בהם יש צורך להציג נתונים המשקפים תמונת מצב בנקודת זמן מסויימת על סמך הפעילויות שהצטברו עד לאותה נקודת זמן.

ביישום זה, הדרישה היא לממש תהליך הבונה ומעדכן טבלת Snapshot על סמך הנתונים בטבלאות המימד והעובדה שכבר הוזנו במחסן הנתונים.

* כהקדמה ליישום – יש לחשוב על **תרחיש עסקי שיצדיק שימוש בטבלת עובדה מסוג Snapshot** לצורך הפקת יישומי BI&A ותמיכה בקבלת החלטות. צוות שבחר ביישום זה יידרש להמחיש שימוש בטבלת Snapshot לצורך הפקה של יישום רלוונטי ומשכנע. יישום טבלת Snapshot שאין לה הצדקה עסקית של ממש, עלול להוביל להפחתת ניקוד.
* בתכנון טבלת Snapshot יש להביא לידי ביטוי מימד זמן אחד (תאריך או שעה, בהתאם לנדרש), קישור לשני מימדים נוספים מאלו הקיימים במחסן הנתונים, ולפחות 2-3 משתני עובדה שעשויים להיות רלוונטיים לתכנון הטבלה והשימוש בה בהמשך.
* את טבלת ה-Snapshot יש להזין מהנתונים שכבר נצברו בטבלת העובדה הפרטנית שיושמה בחלק הקודם, ולא ישירות מנתוני המקור.

במהלך מפגש ההגנה המסכם תידרשו להמחיש באמצעות דוגמא משכנעת את פעולת תהליך ה-ETL שמעדכנים את טבלת ה-Snapshot על סמך הנתונים שנאגרו בטבלת העובדה הפרטנית במחסן הנתונים.

* הההדגמה צריכה לכלול בנייה ראשונית של טבלת Snapshot, על סמך נתונים שכבר קיימים בטבלת העובדה הפרטנית. לאחר מכן – יש להוסיף באופן אינקרמנטלי רשומות חדשות לטבלת העובדה הפרטנית, ולהמחיש כיצד תהליך ה-ETL מעדכן את טבלת ה-Snapshot בהתאם.
* בנוסף להדגמת תהליך ה-ETL, תידרשו בחלק ג' של הפרויקט גם להמחיש שימוש בטבלת ה-Snapshot לבניית יישום רלוונטי למשתמשי הקצה.

**עבור יישום זה, יש להראות בסרטון:**

* הסבר קצר של הרעיון העסקי העומד מאחורי טבלת תמונת המצב ומצדיק את מימושה.
* המחשת פעולת התהליך באמצעות הוספת מספר רשומות לדוגמא לטבלת העובדה הפרטנית, ולאחר מכן להראות שאכן מתבצעת הצבירה הנכונה בטבלת תמונת המצב.

**ג. (10%) בדיקת "יושרת היחס" (RI - Referential Integrity) של נתוני טבלאות העובדה**

בדיקת "יושרת היחס" (RI - Referential) נדרשת במצבים בהם במקור הנתונים אין אכיפה קפדנית של אילוצי מפתח זר, וכתוצאה מכך ברשומות שמתארות פעילויות עסקיות (טרנזקציות) ומיועדות לטבלת העובדה יופיעו ערכי מפתח שאינם קיימים בטבלת המימד. בקורס הוצגו שתי גישות אפשריות לטיפול בכשלי RI:

1. רשומות שהתגלה בהן כשל RI לא יוזנו לטבלת ה-FACT, אלא יופנו לטבלה נפרדת לצורך דיווח שגיאות ותחקור נוסף.
2. לטבלת המימד תתווסף רשומה מלאכותית, שתייצג את הערך הבלתי מוכר (עם קידוד "Unknown"). רשומת הפעולה שבה זוהה הערך הבלתי מוכר תקושר לרשומה המלאכותית, ותוזן לטבלת ה-FACT. כמובן שגם במקרה כזה יש צורך לדווח על המקרים הלא מוכרים שהתגלו, ולאפשר תחקור של הבעיה.

ביישום זה הדרישה היא לבנות תהליכי ETL המבצעים אימות RI בשלב ה-STG, עבור רשומות המיועדות לטבלת העובדה הפרטנית, מול אחת מטבלאות המימד לבחירתכם (למעט מימדי תאריך/שעה).

* היישום צריך לממש בנפרד כל אחת מהגישות האפשריות – הן את האפשרות להפריד את הרשומות הבעייתיות לטבלה נפרדת, והן את האפשרות ליצור רשומת מימד מלאכותית.
* לכל גישה יש לצרף דו"ח מתאים (באמצעות Tableau), שיציג למשתמש את הרשומות הבעייתיות בצורה סדורה, שתאפשר בעתיד תחקור נוסף. במפגש ההגנה תצטרכו להציג את הדוחות האלו.

במהלך מפגש ההגנה המסכם, תידרשו להמחיש באמצעות דוגמא מתאימה הזנה של רשומות עובדה שניתן להבחין בהן בפגמי RI וטיפול נכון של תהליך ה-ETL ברשומות אלו - בגישה הראשונה לעומת הגישה השנייה – בליווי הפקת דו"ח מתאים.

**עבור ישום זה יש להציג סרטון יחיד הממחיש את שתי הגישות:**

* **גישה א':** הוספת רשומת עובדה עם כשל לטבלת העובדה שמציינת Rejected.
* **גישה ב':** הוספת רשומת ממד עם קידוד Unknown”", והוספת רשמות עובדה בהתאם לטבלה הפרטנית.
* שימו לב – במפגש ההגנה תצטרכו להראות בנוסף לסרטון גם אפשרות לתחקור נוח של רשומות עם כשל.