|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| מס' קבוצה | שם הפרוייקט | | | תאריך הגשה |
| 5 | Princess Cruises | | | 30/03/2019 |
| מספרי תעודות הזהות של המגישים | | | | |
| 204504377 | | 311357529 | 203016019 | |

## **(יש להסיר את עמודי ההסבר מהתבנית לפני שמגישים את המטלה)**

**חלק ב' - הקמת תשתית הנתונים**

## **חלק ב' בכללותו מהווה 10% מציון הקורס.** מטרתו היא עיצוב והקמת תשתית הנתונים ע״פ האפיון הראשוני בחלק א׳.

**לאחר השלמת יישום מחסן הנתונים ותהליך ה-ETL - יש לתאם פגישה עם עוזר/ת ההוראה לצורך הצגת התוצרים. את המסמך המלווה יש להגיש אלקטרונית בדוא"ל לעוזר/ת ההוראה, לפני המפגש.**

## **להלן מספר הנחיות כלליות לגבי חלק זה:**

## הפרויקט יבוצע ויוגש בצוותים שנקבעו בחלקים הקודמים של העבודה.

## יש להגיש את כל המטלות במועד על פי ההנחיות שתפורסמנה ב Moodle. הגשה מאוחרת מחייבת קבלת אישור מהמרצה ו/או מעוזר/ת ההוראה.

## 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.

## יש להקפיד על הגשה תמציתית וברורה, בהתאם להגבלה על מספר העמודים בכל חלק. **אי הקפדה על כך עלולה להוביל להורדה בציון**.

## יש להגיש את כל המטלות הכתובות בקובץ מודפס 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.

# (20%) יישום מקורות הנתונים ומחסן הנתונים הטבלאי

## ***(כ- 1-2 עמודים)***

כחלק מהתיקונים אליהם נדרשנו מהחלק הראשון בפרויקט, ביצענו את השינויים הבאים:

1. התיקון הראשון הוא יצירת טבלת Summary אליה יחובר המימד של בסיס הנתונים המשני (GDP) . בחרנו להציג טבלה זו ברמת גרעיניות גבוהה על מנת לאפשר פילוח

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 ב' של הפרויקט.

* יש לכלול במסמך תרשימים טבלאיים מתוקנים.
* לכל מודל – יש לתאר בקצרה, ברשימת נקודות תמציתית, מה השינויים העיקריים שנעשו בו לעומת המודל שהוגש בחלק א'.

באמצאות שרת MySQL

* 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.
* יש לאכלס כל טבלה בנתונים
  + 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 בין ישויות.
  + הנתונים צריכים להיות קריאים, מעניינים ובעלי הגיון עסקי. אם הם הופקו באמצעות סימולציה – אין להשתמש במחולל המייצר נתוני "ג'יבריש" חסרי משמעות. מומלץ גם שלא להפיק את משתני העובדה באופן אקראי לגמרי, אלא לחשוב על תהליך הפקה (למשל, באמצעות הפונקציות הקיימות ב-Excel) שיוביל לאוסף נתונים בעל משמעות עסקית מעניינת (מגמות לאורך זמן, השפעה הדדית, שונות בין קבוצות, וכד')
  + שימו לב, שבעת מילוי הנתונים יש לשמור על האילוצים. למשל; בטבלת הזמנות לא ניתן ליצור הזמנה עם לקוח, שלא קיים בטבלת לקוחות.

בנוסף לכך, יש להקים באמצעות ה- MySQL גם את טבלאות מחסן הנתונים על פי המבנה המתוקן.

# (30%) אפיון תהליכי ה ETL

**סדר הרצת השלבים השונים בתהליך ה-ETL:**

**Data Mirroring (MRR)- -** שלב זה כולל העתקה של כל בסיס הנתונים מהטבלאות המקוריות הנמצאות ב- OLTP, שהוא שלב מקדים לשלב זה, לסביבת העבודה של מחסן הנתונים. שלב זה יתבצע במטרה להמשיך לעבוד על הנתונים העדכניים ביותר מבלי להעמיס על בסיס הנתונים המקורי, וכן מבלי למנוע שיבוש של נתוני המקור. בשלב זה מבנה הטבלאות המקור יישאר ללא שינוי. טרם יצירת הטבלאות נבצע מחיקה של הטבלאות הקודמות, עבור נתוני טבלת העובדה נבצע טעינה אינקרמנטלית, כלומר נטען רק את השינויים שבוצעו בפרק הזמן הוקצב שהוגדר. הבחירה בטעינה אינקרמנטלית נעשית מכיוון שמדובר בכמות רבה של נתונים אותם נידרש להעביר בשלב זה, כאשר רק חלק מסוים אכן השתנה בפרק הזמן המוקצב (delta).

**Dimensions Staging (STG-DIMS)-** שלב זה כולל את ההכנה של נתוני המימדים והעברתם לטבלאות על בסיס סכמת הכוכב ב- DW. שלב זה יתבצע לפי היררכיה- ראשית נעביר טבלאות ברמת היררכיה גבוהה יותר בכדי לא לפגוע בקשרים למפתחות זרים. כמו כן, שלב זה יכיל רק את הנתונים העדכניים של הריצה האחרונה, ובשל כך שלב זה קריטי שכן הוא מאפשר בדיקת נתונים עדכניים ואיתור מוקדם של טעויות ותקלות בתהליך ה-ETL. נבצע מחיקה של הנתונים שכבר הועברו למחסן הנתונים.

**Dimensions Warehousing (DW-DIMS)-** שלב זה כולל עדכון של טבלאות המימדים בDW.  
עדכון זה יתבצע באמצעות הוספת STG DIM למחסן הנתונים בDW באמצעות הפעולה Upsert[[1]](#footnote-1). בשונה משאר השלבים, בשלב זה לא תתבצע מחיקה של נתונים קודמים מאחר וה- DW שומר את היסטוריית הנתונים.

**Fact Staging (STG-FACT)-** שלב זה כולל הכנה של נתוני העובדות והעברתם לטבלאות על פי סכימת הכוכב. שלב זה יתבצע לאחר שיש בידינו את הנתונים של הממדים ב-STG, ולכן נידרש לבצע בדיקה שאכן קיימת תאימות עם טבלאות הממד הנדרשות.

**Referential Integrity (RI)-** שלב זה כולל בדיקת ולידציה ובדיקת שלמות הנתונים בטבלאות ה-Facts אל מול טבלאות המימד. כלומר, מתבצעת בדיקה האם רשומה שנמצאת בטבלת העובדות מכילה ערך שלא קיים בטבלת המימד (מפתח זר שאינו מקושר). נדגיש כי על כל הרשומות בטבלאות העובדה נבצע בדיקה זו. במידה וקיימת סתירה ניתן לפתור באחת משתי הדרכים הבאות:

1. ניתן להעביר את הרשומה לטבלה זמנית, ולאמת את קיום הערך של הרשומה פעם נוספת בפעם הבאה.
2. **לחכות לתשובה מדן המרצה**

נבצע הוספה של אובייקט זה לטבלת הממד ונקשר אותו לרשומה שלו בטבלת העובדות. בשלב זה נרוץ על המידע המעודכן ולא על כל היסטורית העובדות.

**Fact Warehousing (DW-FACT)-** שלב זה כולל עדכון של טבלאות העובדות בDW. הוספת נתונים חדשים תתבצע באמצעות הפעולה insert באופן אינקרמנטלי (מסיבה דומה לזו שהוסברה בשלבים קודמים).

**התלויות הקיימות בין שלבי התהליך:**

על מנת שתהליך ה-ETL באופן אופטימלי, כלומר כזה שיאפשר לנו לבצע את התהליך בצורה היעילה, האיכותית והמהירה ביותר, חשוב שהשלבים הנ"ל יתבצעו בסדר הרשום. קיימות תלויות בין השלבים השונים שכן כל שלב מתבסס על קודמו, ולכן לא ניתן לבצע שלב מתקדם לפני שביצענו את השלבים שקדמו לו. תחילה מתבצע תהליך מלא של MRR עבור כל הטבלאות ומשם מתחיל STG והכנסה של טבלאות הממד למחסן הנתונים ורק לאחר מכן מתבצע תהליך דומה עבור טבלאות הFACT. כמו כן, קיימת תלות בין טעינת טבלאות המימדים לבין טעינת טבלאות העובדות. יש צורך לטעון קודם רק את הממדים ורק לאחר מכן את העובדות, מכיוון שטבלאות העובדות מצביעות על טבלאות המימדים (ולמעשה השדות המזהים בטבלאות העובדה מבוססות על טבלאות המימדים.

**במערכת האמת השלבים הבאים צפויים לקחת זמן רב במיוחד:**

במערכת האמת סביר שתהיה כמות גדולה של רשומות ולכן ישנם מספר שלבים אשר צפויים לקחת זמן רב:

1. שלב הSTG צפוי לקחת זמן רב מכיוון שבשלב זה מתבצעות פעולות join. פעולות אלו יידרשו זמן רב מכיוון שהטבלאות יהיו בגודל מלא ויידרש ביצוע אגרגציות ופעולות חישוביות שעשוית להשפיע ולהאריך משמעותית את משך תהליך הSTG.
2. שלב הRI- בדיקות אימות הנתונים עלולות לקחת זמן רב במידה וישנן כמויות גדולות של נתונים.
3. שלב הטעינה ל-DW יכול לקחת זמן רב במידה וישנן כמויות גדולות של נתונים.

שלב מוקדם שעשוי לעכב את כלל התהליך הוא שלב טעינת בסיס הנתונים כשלב מקדים ל-ETL, ב-OLTP. אמנם שלב זה לא נעשה ב-ETL, אך זהו שלב מקדים שעשוי לעכב את התהליך כולו.

**שלבי הETL- שעשויים להיכשל ודרכי ההתמודדות:**

* כשל אפשרי: ייתכן כשל בתהליך עדכון הנתונים אשר עשוי לפגוע בפעילות השוטפת של בסיס הנתונים התפעולי. דבר זה עלול להיגרם עקב עומס על המערכות.

פתרון: בכדי להתמודד עם בעיה זו, נרצה לבצע את העתקות הנתונים ממערכת המקור לMRR בזמנים בהם יש פחות עומס על המערכת בהתאם למאפייני ושעות העבודה עם בסיס הנתונים (בשעות הלילה, סופי שבוע, או בזמנים בהם העבודה על בסיסי הנתונים מועטה יותר). בנוסף, ניתן לדאוג לתיאום בין הגורמים המוסמכים לעבוד על בסיס הנתונים ולאיזון העבודה ביניהם.

* כשל: נתונים חסרים או לא תקינים בטבלאות התפעוליות שלא עברו בקרה לפני הכנסתם אל סביבת העבודה של מחסן הנתונים עלולים להוות בעיה של איכות נתונים. פתרון: קשה מאוד להתמודד עם בעיה זו בשלב זה. ישנם כלים שיכולים להטיב עם הנתונים, ולמנוע תקלות אפשריות בטעינת הנתונים (תהליך Data Cleansing).
* כשל: בעת חיפוש מפתחות זרים בין טבלאות ה-Fact והמימדים ניתקל ברשומות Null ובשגיאות אחרות.

פתרון: בעת חיפוש עבור טבלאות הממד ישנה חשיבות רבה לסדר טעינת הטבלאות התפעוליות- יש לטעון אותם בסדר מתאים כך ששדה של מפתח זר יפנה אך ורק לשדה של טבלה שכבר עודכנה. אחרת, לשם כך שלב ה-RI הוא שלב קריטי, שכאמור מאפשר בדיקות תאימות שימנעו פנייה אפשרית לרשומות שלא התעדכנו או נוספו. פתרון שכזה יכלול למשל העברת רשומה בעלת ערך Null בטבלאות העובדה לטבלת "דחויים" זמנית, ולבדוק בהמשך את סיבת הבעיה שנוצרה.

**STT**

# (50%) מימוש תהליכי ה-ETL ואכלוס מחסן הנתונים

***(כ- 1-2 עמודים)***

יש לפתח את **מנגנון ה-ETL** שימיר את הנתונים משני המקורות, בהתאם לאפיון ה STT שאפיינתם בחלק א'. מנגנון ה-ETL צריך לבצע את כל השלבים הדרושים בתהליך, על פי המתודולוגיה אשר נלמדה בכיתה - שאיבת הנתונים ממקורות הנתונים, המרתם למבנה הדרוש, ואכלוס הנתונים המומרים בטבלאות המתאימות.

**מימוש ה-ETL יתבצע באמצעות כלי ה- (Talend) Data Integration, המותקן במעבדות ההוראה וכן זמין להורדה למחשבכם האישי.**

* 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 תהליך ה-ETL על שלביו השונים. יש לכלול צילומי מסך של הסכמות הגרפיות הרלוונטיות במסמך זה.
* במהלך המפגש עם הבודק/ת תצטרכו להציג את תהליך ה-ETL שבניתם. **כמו כן, יש להציג מפגש גם את ה- SQL Script הבונה את מחסן הנתונים (סדרת פעולות CREATE TABLE)**
  + **שימו לב:** הינכם נדרשים לממש לפחות מימד אחד כמשתנה לאט מסוג 2!
* **מימד התאריך והשעה-** אלו מימדים חשובים ביותר אשר צריכים להופיע בכל מחסן נתונים, לנוחיותכם בתאר הקורס קוד ליצירת המימדים (ניתן לייצרם ישירות למחסן הנתונים), כמובן שניתן להשתמש בכל קוד אחר שתראו לנכון.

1. עדכון רשומות קיימות או הוספת רשומה שאינה קיימת. [↑](#footnote-ref-1)