אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev
המחלקה להנדסת מערכות תוכנה ומידע

מעבדה לאיסוף וניהול נתונים

עבודה מספר 2 ב- SSIS

מטרת התרגיל

בתרגיל זה תפתחו תהליכי ETL ב- SSIS המממשים איסוף מידע ממספר מקורות נתונים דינמיים, העברת הנתונים הרלוונטיים לאזור Staging דינמי, ביצוע טרנספורמציות ודיווח על שגיאות.

הנחיות הגשה

- יש להגיש את כל המטלות במועד על פי ההנחיות שתפורסמנה ב Moodle.
- הגשה באיחור תגרור הורדת ניקוד.
- את המטלה יש להגיש לתיבת ההגשה ב- Moodle. קובץ הפתרון יהיה קובץ מכווץ מסוג zip ששמו מורכב ממספרי תעודת הזהות של המגישים, כאשר ביניהם ישנו קו תחתון.(לדוגמה: ID1_ID2.zip)
- קובץ ההגשה יכלול את הקבצים הבאים:
  - קבצי SSIS שיצרתם (שיוכלו לשמש להרצה לשם בדיקת הפתרון).
  - קובץ של בסיס הנתונים שנוצר בתום ההרצה.
  - קובץ דו"ח שגיאות ששמו Errors.csv שנוצר בתום ההרצה.
  - קובץ Word שמכיל תרשימים מה– SSIS של ControlFlow ו- DataFlow.
  - הוראות הרצה מיוחדות (רק במידש ונדרש) בקובץ Word.

משימות התרגיל

הפתרון שתספקו אמור לבצע את הפעולות הבאות:

1. צרו מתחת לכונן :C את התיקייה Staging ומתחתיה שלוש תיקיות מקבילות ששמן:
   SharingFolder, ProcessFolder, ArchivedFolder
2. הניחו שבכל יום מועתק לתיקייה SharingFolder קובץ נתונים חדש אשר מכווץ בפורמט ZIP ששמו מכיל את המילה data ואחריה את התאריך של היום בפורמט yyyymmdd
   לדוגמה: בתאריך 25/3/2022 יועתק לתיקייה זו קובץ נתונים ששמו: data20220325.zip
3. פתחו את קובץ הנתונים המכווץ והעתיקו את הקבצים שבתוכו לתיקייה ששמה: ProcessFolder
4. קבצי ה- CSV שנפתחו מכילים ארבע יישויות נתונים שונות (Employee, Hotel, Product, Supplier) ובצמוד אליהם התאריך של היום בפורמט yyyymmdd
   לדוגמה: בתאריך 25/3/2022 יועתק לתיקייה זו ארבעה קבצי CSV ששמם:
   Employee20220325.csv , Hotel20220325.csv , Product 20220325.csv , Supplier 20220325.csv
5. עליכם להעתיק את תוכן קבצי ה- CSV לטבלאות בבסיס הנתונים SQL-Server כאשר כל הרשומות של קובץ מסוים יתווספו לטבלה ששמה הינו כשמו של הקובץ מלבד התאריך.
   לדוגמה: תוכן הקובץ Hotel20220325.csv יתווסף לטבלה ששמה: Hotel
6. שימו לב שבעת טעינת נתוני הקבצים לטבלאות יש לטפל גם במקרים הבאים:

- אם יש נתונים בקבצים שמופיעים בתוך גרשיים (כמו למשל בקובץ Hotel20220325.csv) אז העתיקו את הנתונים לטבלאות ב- SQL-Server ללא הגרשיים.
- לאחר העתקת שדות הקובץ Employee לטבלה ב- SQL-Server , יש להעתיק עבור כל רשומת עובד במקום השדה GeographyID את כל שדות הקובץ של החברה שבה הוא עובד מתוך הקובץ Company.txt , כך שבסופו של דבר

כל רשומה בטבלת Employee ב- SQL-Server תכיל את כל השדות מטבלת העובד מקובץ ה- CSV ובנוסף כל השדות מטבלת החברה מקובץ הטקס, מלבד השדה GeographyID . יש לקשר בין העובד לחברה שלו באמצעות השדה GeographyID אך אין לעשות זאת ע"י העתקת נתוני קובץ הטקסט לטבלה ב- SQL-Server ואין לבצע JOIN בין שני טבלאות בבסיס הנתונים על מנת לפתור זאת אלא שיש לפתור זאת ע"י JOIN בין שני מקורות מידע שונים של קובץ וטבלה.

- הוסיפו בכל טבלה לכל רשומה שדה שמכיל את תאריך ושעת יצירת הרשומה.

7. לאחר סיום העתקת הנתונים מכל קובץ לבסיס הנתונים, מחקו את את הקובץ מתיקיית ProcessFolder
8. בסיום המעבר על כל הקבצים, העבירו את קובץ הנתונים המכווץ (לדוגמה: data20220325.zip) לתיקייה ששמה: ArchivedFolder
9. ה- Package אמור להציג הודעות שגיאה לאורך כל התהליך, לדוגמה:
   - אם אין קבצים בתיקייה SharingFolder
   - אם שמות הקבצים אינם מכילים את התאריך של היום
   - תקלה בטעינת נתוני הקבצים לטבלאות בבסיס הנתונים
   - ועוד...
10. שמרו לוג של כל השגיאות בטבלת Log שמכיל את פרטי כל השגיאות, באיזה שלב בתהליך הן קרו ואת זמן התרחשותן.
11. לשם בדיקת הפתרון, השתמשו בקובץ ה- ZIP המצורף ועדכנו את התאריכים בשמות הקבצים של קובץ ה-ZIP וגם של כל הקבצים בתוכו.
12. לאחר סיום כתיבת ובדיקת הפתרון, יש להריץ את הפתרון פעמיים: פעם אחת כאשר התאריך בכל שמות הקבצים שסופקו הינו התאריך של יום ההרצה ואילו בפעם השניה התאריך שונה מתאריך יום ההרצה.
13. הפיקו דו"ח בסוף ההרצה על כל השגיאות שתיעדתם ע"י יצירת קובץ ששמו Errors.csv שמכיל את כל פרטי טבלת Log.
14. יש לתמוך בשמות דינמיים של ארבעת התיקיות ששמן: Staging, SharingFolder, ProcessFolder, ArchivedFolder
    כלומר, אם המשתמש משנה את שמות התיקיות הללו אז ה- Package ימשיך לעבוד כפי שנדרש וללא תקלות.
    יש לתמוך במיקום דינמי של תיקיית SharingFolder, כלומר אם המשתמש משנה את המיקום של תיקייה זו אז ה- Package ימשיך לעבוד כפי שנדרש וללא תקלות.

<u>הארות והבהרות:</u>

- יש לממש את כל התהליך שלעיל באמצעות Package (או יותר) של SSIS.
- שלושת התיקיות ששמן SharingFolder, ProcessFolder, ArchivedFolder אמורות להיות ממוקמות מתחת לתיקיית Staging, גם אם השם של כל אחד מארבעת התיקיות ישתנה.
- שימו לב להשתמש בשמות משמעותיים במשימות (Tasks) ובאנוטציות על מנת להסביר תהליכים מורכבים ב- SSIS.

<u>בהצלחה</u>