הפקולטה להנדסה
ע"ש איבי ואלדר פליישמן
אוניברסיטת תל אביב

מחסני נתונים - דו"ח תרגיל 2

חברת Stratton Oakmont

JUNE 27, 2023

דניאל וולקוביץ, ת.ז : 207257668
אלעד פרומן, ת.ז : 207231143

תוכן עניינים

עיבוד וניקוי הנתונים

ניתוח ראשוני

המידע שהתקבל חולק לטבלאות מטא-דאטא, ולטבלאות tbl – אשר יחדיו מרכיבות את מחסן הנתונים, כאשר החומר הגולמי כמובן נשמר על שבעת שהחברה תרצה בכך, היא תמיד תוכל לשחזר אותם.

טבלאות מטא-דאטא

על מנת לבצע סטנדרטיזציה לנתונים וכדי לשמור על הסדר, הזנו טבלאות מטא-דאטא אשר מכילות תרגום בין נתונים שהתקבלו מהטבלאות השונות, לבין קודים מספריים אחידים וייחודיים. כך, יצרנו את הטבלאות הבאות:

1. meta.stock_type_code – טבלא אשר ממפה בין קוד המנייה לסוג המנייה (האם היא זולה או יקרה)
2. meta.stocks_code_to_name – טבלא אשר ממפה בין קוד המנייה לבין השם שלה
3. meta.valid_emails – טבלא אשר מכילה את כלל כתובות המייל הוולידיות אשר ניתן להזין למחסן הנתונים.

צעדי ניקוי ועיבוד

להלן הטבלאות אשר בחרנו לייצר מתוך המידע הגולמי ואשר מרכיבות את מחסן הנתונים:

tbl.brokers

טבלאת מימד אשר מכילה מידע על כלל הברוקרים אשר עובדים בחברה, כאשר בטבלא זו בוצעו שינויים קלים בלבד והיא דומה לטבלא הגולמית אשר התקבלה.

tbl.exchangerates

טבלאת מימד אשר מכילה את שערי החליפין בכל יום, כאשר בטבלא זו בוצעו שינויים קלים בלבד והיא דומה לטבלא הגולמית אשר התקבלה.

tbl.transactions

טבלאת עובדה אשר מכילה את הטרנזקציות המוצלחות בלבד שהתרחשו בחברה. בטבלא זו נעשו צעדי סידור וניקוי רבים, כאשר הטבלא הסופית מכיל את העמודות הבאות: מזהה הברוקר, תאריך, מזהה המנייה, מזהה המשקיע, האם מדובר במכירה או בקניית מניות (בינארי) ומהו שווי הטרנזקציה בדולרים.

tbl.failed_transactions

טבלאת עובדה אשר מכילה את הטרנזקציות אשר נכשלו מסיבה כזו או אחרת. כאשר את עמודת value שהגיע מן הטבלא הגולמית הוחלט להסיר מכיוון שהכילה נימוקים לא רלוונטיים ולא מועילים לכישלון העסקה.

tbl.stock_spots

טבלאת מימד אשר מכילה את ערכי המניות בכל יום ויום, כאשר בטבלא זו בוצעו שינויים קלים בלבד והיא דומה לטבלא הגולמית אשר התקבלה.

tbl.stocks

טבלאת מימד אשר מכילה את מניות המסחר הקיימות בחברה, הטבלא שונתה בהתאם לטבלאות המטא שיצרנו, ומכילה קוד זיהוי ייחודי לכל מנייה, וקוד ייחודי נוסף לכל סוג מנייה.

tbl.investors ו- tbl.state

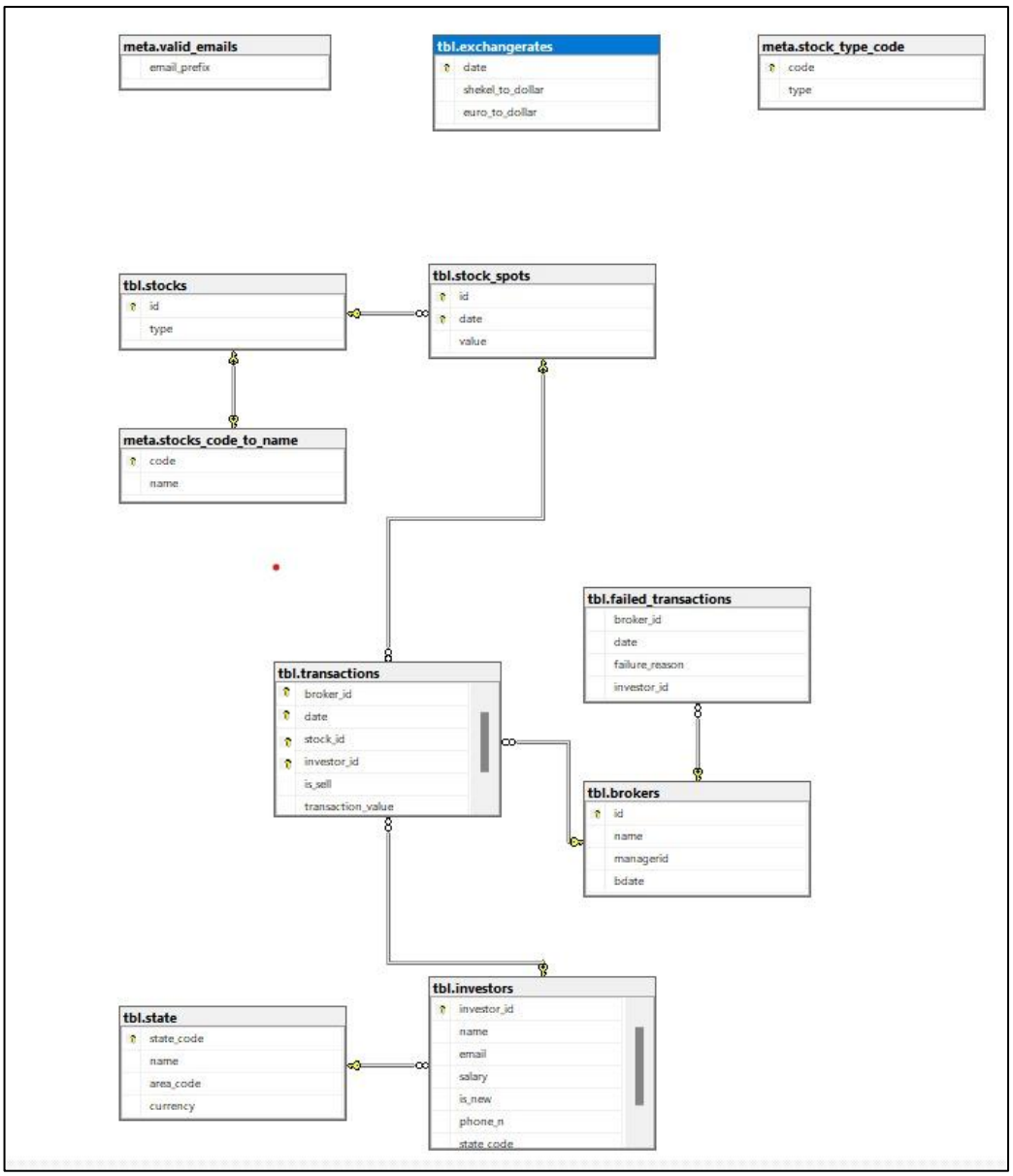
מן המידע הגומי התקבלו טבלאות אשר מכילות מידע על משקיע עבר ועל משקיעים חדשים.
כחלק מתהליך העיבוד הוחלט לחלק את הטבלאות הללו לשני טבלאות שונות:

- טבלת משקיעים אשר מכילה פרטים על המשקיע, כולל קוד מדינה.
- טבלת מדינה, שמכילה את הפרטים על המדינה- כולל קידומת, שם, קוד מדינה ומטבע.

כך ביצענו תהליך של סטנדרטיזציה לשדה מספר הטלפון, כאשר קידומת המדינה שמורה בשדה נפרד מאשר שאר מספר הטלפון. כאשר מספר הטלפון ללא קידומת נמצא בטבלת המשקיעים, וקידומת המדינה המתאימה לקוד המדינה נמצאת בטבלת המדינה.
בנוסף, הוספנו טריגר שמוודא את פורמט המייל של המשקיע- ומאפשר הוספת כתובות מייל חוקיות נוספות, אך מעלה שגיאה כאשר כתובת מייל איננה חוקית.

תרשים הטבלאות

להלן תרשים הטבלאות, אשר מתאר ה-constraints שהוספנו לכל טבלא.



מרכולי הנתונים

את מרכולי הנתונים בחרנו להציג כ-VIEWS. מכיוון למעשה מדובר על טבלאות אשר משקפות את הנתונים הקיימים במחסן הנתונים בצורה "שונה" בלבד, ואין מדובר בנתונים חדשים.

מרכול נתונים למנהלי צוותים

לכל ראש צוות מרכול ייעודי עבורו, כאשר כל מרכול מכיל את השדות הבאים:

1. מזהה ברוקר.
2. שם ברוקר.
3. כמות המשקיעים עימם הברוקר נמצא בקשר.
4. שווי מכירת המניות הכולל שהברוקר ביצע.
5. שווי קניית המניות הכולל שהברוקר ביצע.
6. שווי טרנזקציה ממוצעת שהברוקר מבצע.

7. סכום כלל הטרנזצקיות שהברוקר ביצע.
8. שווי טרנזקציית המכירה הגדולה ביותר.
9. שווי טרנזקציית הקנייה הגדולה ביותר.
10. ממוצע מספר הטרנזקציות שמבצע הברוקר ביום.
11. משך הזמן בו הברוקר עובד בחברה.

כמוכן המרכול מכיל את ה-KPI "המורכבים יותר" הבאים:

12. מדד הצלחת המכירות – אחוז הטרנזקציות המוצלחות שהברוקר ביצע כאחוז מתוך מספר הטרנזקציות הכללי שהוא ביצע.
13. מדד retention – מדד אשר מודד עד כמה מצליח הברוקר לשמור על קשר עם הלקוחות הקיימים שלו.

מרכול נתונים למנכ"ל החברה

למנכ"ל מרכול ייעודי עבורו, כאשר המרכול מכיל את השדות הבאים:

1. שם ראש הצוות.
2. כמות חברי הצוות.
3. טרנזקציה ממוצעת בצוות.
4. הטרזנקציה הגבוהה ביותר בצוות.
5. הטרנזקציה הנמוכה ביותר בצוות.
6. מספר הטרנזקציות בחלוקה למספר הברוקרים.

מרכול נתונים להנהלת חשבונות

המרכול הוזן כ-view בהתאם להנחיות.

מרכול נתונים לסמנכ"ל כספים

המרכול הוזן כ-view בהתאם להנחיות.

תהליך pipe-line

יצרנו pipe-line אשר משמש כשלב ה-TL בתהליך ETL אפשרי. תהליך זה פועל באופן תקין כאשר המשתמש מנסה לייצר לקוחות חדשים ע"י אספקת טבלת לקוחות חדשים, אשר ניתנת בפורמט זהה לפורמט הנתון בטבלה הגולמית אשר ניתנה בתרגיל (NEW INVESTORS).
במסגרת תהליך זה עשינו שימוש בשלוש פונקציות ובפרוצדורה אחת.

ראשית, יצרנו פונקצייה אשר מבצעת את שלב הטרנספורמציה, היא מנרמלת את המידע, וממירה אותו לפורמט של טבלת המשקיעים אשר יצרנו במחסן הנתונים שלנו (קרי, תהליך הנרמול כולל הפרדה לתתי טבלאות שמכילות מידע על המדינה והמטבע וכו').

שנית, הקמנו פרוצדורה שמורה שטוענת את הטבלה שעברה נירמול, אל תוך מסד הנתונים. כאשר הסיבה לא לבצע את כלל השלבים בפונקציה אחת היא שנוכל לבצע בדיקות של תקינות הקלט המנורמל לפני שנטען אותו לתוך מחסן הנתונים אשר מכיל את המידע הרגיש שלנו.

את בדיקת התקינות המדוברות מימשנו באמצעות שתי פונקציות נוספות. הראשונה בודקת שהטבלה לאחר הטרנספורמציה היא בפורמט הרצוי, זאת על ידי יצירת עמודה שמציינת לכל רשומה האם היא ולידית או לא. ואילו הפונקצייה השנייה בודקת האם כל הרשומות ולידיות, ומסירה את עמודת הולידציה במידה שזהו אכן המצב.

כך למעשה, תהליך ה-TL המלא מתנהל באופן הבא:

1. נכנסת טבלה של משקיעים חדשים.
2. הטבלא עוברת סטנדרטיזציה.
3. נעשית בדיקה של הרשומות לאחר סטנדרטיזציה.
4. אם כל העמודות ולידיות, נסיר את עמודת הולידציה ונעבור לטעינת המידע.
   אחרת, לא נוכל להסיר את העמודה, וטעינת המידע תיכשל - כך נוכל לבדוק את המידע ולתקן אותו כנדרש.
5. נטען את המידע למחסן הנתונים במידה שהמידע תוקן כראוי.

פונקציות

עפ״י הדרישות, נדרשנו לכתוב 3 פונקציות, וכולן שימשו אותנו בתהליך ה-pipe line שתואר לעיל.

פרוצדורות

בתהליך ה-pipe line שתואר לעיל עשינו שימוש בפרוצדורה אחת, ועל כן הוספנו פרוצדורה נוספת. הפרוצדורה מאפשרת להזין מדינה חדשה אל תוך טבלאת state, כך במידה שהחברה תעסוק עם לקוחות ממדינות זרות, היא תוכל להוסיף את המדינות הללו למאגר בקלות ובנוחות. הפרוצדורה מוודאה כי כלל הנתונים תקינים וכי ניתן להוסיף את הנתונים אל המאגר באופן תקין, במידה שלא – הפרוצדורה תתריע על כך בהודעת שגיאה.

טריגרים

PreventFutureDates

טריגר אשר מונע הכנסת תאריך עתידי אל טבלאת ה-stock_spots.

trg_check_valid_email

טריגר אשר מונע הכנסת כתובת אימייל שלא הוגדרה כוולידית אל תוך מחסן הנתונים.

check_stock_type

טריגר אשר מונע הכנסת מניות ללא שדה type.