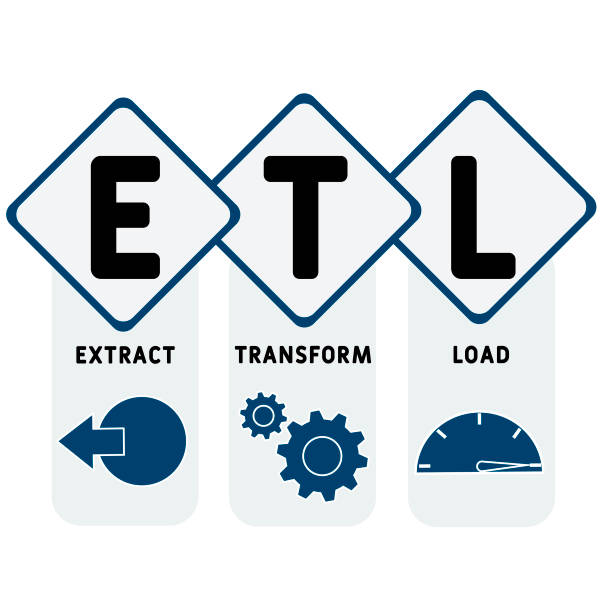
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **מס' קבוצה** | **שם הפרוייקט** | | | **תאריך הגשה** |
| 14 | <https://www.nixon.com/us/en> | | | 17/09/2022 |
| **מספרי תעודות הזהות של המגישים** | | | | |
| 316082015 | | 315633776 | 208461822 | |

  
  
“Getting information off the internet is like taking a drink from a firehose.”  
 – [Mitchell Kap](https://www.computerhope.com/people/mitchell_kapor.htm)or  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
**חלק ב' – הכנת תשתית הנתונים**

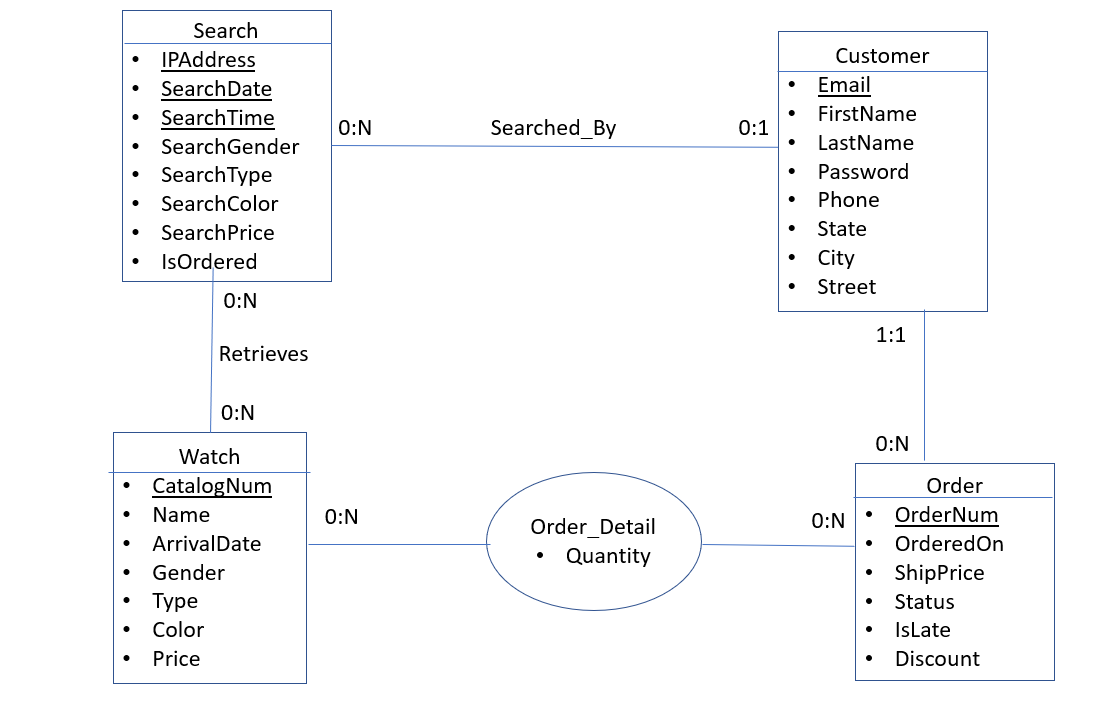
# (20%) יישום מקורות הנתונים ומחסן הנתונים הטבלאי

**עדכון סכמות הנתונים**

# ניתוח מקורות הנתונים התפעוליים

**מקור הנתונים הראשי Nixon – Database:**

**תרשים ERD מתוקן:**

****

**מודל טבלאי:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **הטבלה** | **שם השדה** | **סוג הנתונים** | **סולם המדידה** | **תיאור השדה** | | |
| **CUSTOMERS** | | | Email | Varchar(40) | שמי | מזהה הלקוח |
| FirstNme | Varchar(20) | שמי | שם פרטי של הלקוח |
| LastName | Varchar(20) | שמי | שם משפחה של הלקוח |
| Password | Varchar(20) | שמי | הסיסמה של הלקוח |
| Phone | Varchar(20) | שמי | מספר הטלפון של הלקוח |
| State | Varchar(20) | שמי | המדינה בה מתגורר הלקוח |
| City | Varchar(20) | שמי | העיר בה מתגורר הלקוח |
| Street | Varchar(20) | שמי | הרחוב בו מתגורר הלקוח |
| **ORDERS** | | | OrderNum | Int | שמי | מזהה ההזמנה |
| OrderedOn | Date | רווח | התאריך בו בוצעה ההזמנה |
| Satus | Varchar(40) | שמי | סטטוס ההזמנה |
| IsLate | Boolean | שמי | האם ההזמנה הגיעה בזמן או לא |
| Email (CUSTOMERS) | Varchar(40) | שמי | מזהה הלקוח של ההזמנה |
| ShipPrice | Money | יחס | מחיר המשלוח |
| Discount | Double | יחס | ההנחה שניתנה עבור ההזמנה |
| **WATCHES** | | | CatalogNum | Int | שמי | מזהה השעון |
| Name | Varchar(20) | שמי | שם הדגם |
| ArivalDate | Date | רווח | תאריך הגעת דגם השעון |
| Gender | Varchar(1) | שמי | המגדר עבורו מיועד השעון |
| Type | Varchar(20) | שמי | סוג השעון |
| Color | Varchar(20) | שמי | צבע השעון |
| Price | Money | יחס | מחיר השעון |
| **SEARCHES** | | | IPAddress | Varchar(15) | שמי | מזהה כתובת המחשב ממנו בוצע החיפוש |
| SearchDate | Datet | רווח | מזהה תאריך החיפוש |
| SearchTime | Time | רווח | מזהה שעת החיפוש |
| SearchGender | Varchar(1) | שמי | חיפוש על פי מגדר |
| SearchType | Varchar(20) | שמי | חיפוש לפי סוג השעון |
| SearchColor | Varchar(20) | שמי | חיפוש על פי צבע |
| SearchPrice | Money | יחס | חיפוש על פי מחיר |
| IsOrdered | Boolean | שמי | האם החיפוש הסתיים בהזמנה או לא |
| **ORDER\_DETAILS** | | | OrderNum (ORDERS) | Int | שמי | מזהה ההזמנה |
| CatNum (WATCHES) | Int | שמי | מזהה השעון |
| Quantity | Int | יחס | מספר יחידות המוצר שנכללו |
| **RETRIEVINGS** | | | IPAddress (SEARCHES) | Varchar(15) | שמי | מזהה כתובת המחשב ממנו בוצע החיפוש |
| SearchDate(SEARCHES) | Datet | רווח | מזהה תאריך החיפוש |
| SearchTime (SEARCHES) | Time | רווח | מזהה שעת החיפוש |
| CatalogNum (WATCHES) | Int | שמי | מזהה השעון |

**מקור הנתונים המשני BLS Yearly Unemployment Rate:  
תרשים מתוקן:**

A picture containing application

Description automatically generated

**מודל טבלאי:**

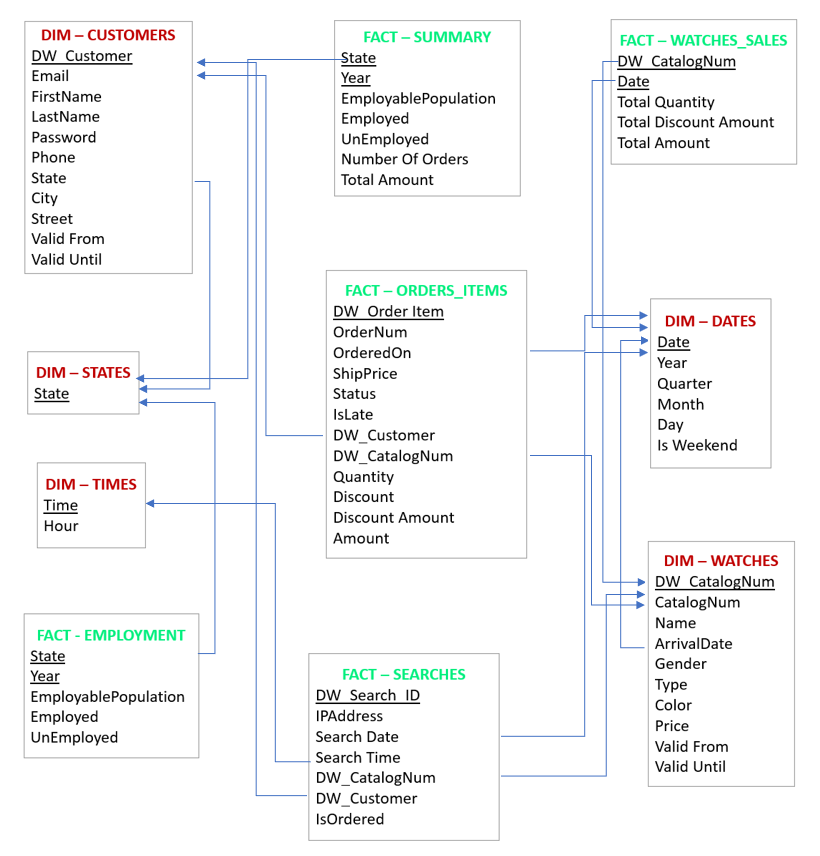
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **הטבלה** | **שם השדה** | **סוג הנתונים** | **סולם המדידה** | **תיאור השדה** |
| **EMPLOYMENT\_STATES** | State | Varchar(20) | שמי | מזהה המדינה |
| Year | Int | רווח | השנה בה נמדד מצב התעסוקה |
| EmployablePopulation | Int | יחס | מספר האזרחים במדינה הכשירים לעבודה |
| Employed | Int | יחס | מספר המועסקים |
| UnEmployed | Int | יחס | מספר המובטלים |

**עיקרי הישוניים שנעשו - מקורות הנתונים התפעוליים:**

# הפכנו גם את השדה שנה למזהה בטבלת State\_Employment

* פיצלנו את השדה SearchDT לשני שדות נפרדים Search Date ו Search Time
* העברנו את השדה Discount מטבלת ORDER\_DETAILS אל טבלת ORDERS כך שההנחה תהיה פר הזמנה ולא פר מוצר בהזמנה.

# תכנון סכמת כוכב



**תיאור מחסן הנתונים**

|  |  |  |  |
| --- | --- | --- | --- |
| **הטבלה** | **שם השדה** | **סוג הנתונים** | **תיאור השדה** |
| **D\_CUSTOMERS**  טבלת מימד משתנה מסוג 2, המכילה את רשימות הלקוחות, תוך שמירת גרסאות | DW\_Customer | Int | מזהה לקוח  מפתח עזר, כל ערך משקף גרסת לקוח בעלת טווח תאריכים מוגדר |
| Email | Varchar(40) | מזהה הלקוח המקורי |
| FirstName | Varchar(20) | שם פרטי של הלקוח |
| LastName | Varchar(20) | שם משפחה של הלקוח |
| Password | Varchar(20) | הסיסמה של הלקוח  שדה זה עשוי להשתנות. |
| Phone | Varchar(20) | מספר הטלפון של הלקוח  שדה זה עשוי להשתנות. |
| State (D\_STATES) | Varchar(20) | המדינה בה מתגורר הלקוח (מפתח זר)  **שדה זה משקף הירארכיה על פיה ניתן לפלח את הלקוחות לפי מדינות.**  **מדינת מגורים עשוייה להשתנות לעיתים, ומכאן הצורך בשמירת גרסאות** |
| City | Varchar(20) | העיר בה מתגורר הלקוח  שדה זה משקף הירארכיה על פיה ניתן לפלח את הלקוחות  עיר המגורים עשוייה להשתנות לעיתים, ומכאן הצורך בשמירת גרסאות |
| Street | Varchar(20) | הרחוב בו מתגורר הלקוח |
| Valid From | Date | תאריך תחילת תוקף הגרסה |
| Valid Until | Date | תאריך סיום תוקף הגרסה |
| **D\_WATCHES**  טבלת מימד משתנה לאט מסוג 2, המכילה את רשימת השעונים, תוך שמירת גרסאות | DW\_CatalogNum | Int | מזהה השעון  מפתח עזר, כל ערך משקף גרסת שעון בעלת טווח תאריכים מוגדר |
| CatalogNum | Int | מזהה השעון המקורי |
| Name | Varchar(20) | שם הדגם |
| ArrivalDate (D\_DATES) | Date | תאריך הגעת דגם השעון (מפתח זר) |
| Gender | Varchar(1) | המגדר עבורו מיועד השעון |
| Type | Varchar(20) | סוג השעון |
| Color | Varchar(20) | צבע השעון |
| Price | Money | מחיר השעון  **שדה זה עשוי להשתנות ומכאן הצורך בשמירת גרסאות** |
| Valid From | Date | תאריך תחילת תוקף הגרסה |
| Valid Until | Date | תאריך סיום תוקף הגרסה |
| **F\_ORDERS\_ITEMS**  טבלת עובדה פרטנית, המכילה את פירוט המוצרים הכלולים בכל הזמנה | DW\_Order\_Item | Int | מזהה פריט מוצר בהזמנה  מפתח עזר, משקף צירוף ייחודי של מזהה הזמנה ומזהה מוצר |
| OrderNum | Int | מזהה ההזמנה המקורי |
| OrderedOn (D\_DATES) | Date | תאריך ביצוע ההזמנה (מפתח זר) |
| ShipPrice | Money | מחיר המשלוח,  משתנה גולמי לאחר ביצוע העמסה |
| Status | Varchar(40) | סטטוס ההזמנה |
| IsLate | Boolean | האם ההזמנה הגיעה בזמן או לא |
| DW\_Customerl (D\_CUSTOMERS) | Int | מזהה הלקוח (מפתח זר) |
| DW\_CatalogNum (D\_WATCHES) | Int | מזהה המוצר (מפתח זר) |
| Quantity | Int | מספר יחידות המוצר, משתנה גולמי |
| Discount | Double | אחוז ההנחה שניתן בעבור פריטים מסוימים בהזמנה |
| Didcount Amount | Money | סכום ההנחה  שדה שמחושב ע"י מכפלת מספר יחידות במחיר ליחידה ובאחוז ההנחה למוצר |
| Amount | Money | סכום הקנייה  שדה שמחושב ע"י מכפלת מספר יחידות במחיר ליחידה פחות סכום ההנחה |
| **D\_STATES**  טבלת מימד משתנה לאט מסוג 1, המכילה את הנתונים הדמוגרפיים העדכניים של 51 המדינות בארצות הברית | State | Varchar(20) | שם המדינה (מזהה) |
| **D\_DATES**  טבלת מימד זמן, המכילה מאפייני תאריך (משתנה לאט מסוג 0) | Date | Date | מזהה תאריך |
| Year | Int | שנה |
| Quarter | Int | רבעון |
| Month | Int | חודש |
| Day | Int | יום בחודש |
| Is Weekwnd | Boolean | מתאר האם התאריך בסוף שבוע. |
| **F\_WATCHES\_SALES**  טבלת עובדה סיכומית, מסכמת את ההזמנות לפי דגם שעון ותאריך | DW\_CatalogNum  (D\_WATCHES) | Int | מזהה השעון (מפתח זר) |
| Date  (D\_DATES) | Date | תאריך (מפתח זר) |
| Total Quantity | Int | כמות פריטים כוללת  שדה שמחושב ע"י SUM על Quantity |
| Total Discount Amount | Money | סכום הנחות כולל  שדה שמחושב ע"י SUM על שדה Discount Amount |
| Total Amount | Money | סכום קניות כולל  שדה שמחושב ע"י SUM על שדה Amount |
| **D\_TIMES**  טבלת מימד זמן, המכילה מאפייני שעה (משתנה לאט מסוג 0) | Time | Time | מזהה שעה |
| Hour | Int | שעה |
| **F\_SEARCHES**  טבלת עובדה פרטנית, מכילה את הפרטים אודות חיפוש המוצרים באתר | DW\_Search\_ID | Int | מזהה החיפוש (מפתח עזר) |
| IPAddress | Varchar(15) | כתובת המחשב ממנו בוצע החיפוש |
| Search Date (D\_DATES) | Date | תאריך החיפוש (מפתח זר) |
| Search Time (D\_TIMES) | Time | מזהה שעת החיפוש |
| DW\_CatalogNum (D\_WATCHES) | Int | מזהה השעון (מפתח זר) |
| DW\_Customer (D\_CUSTOMERS) | Varchar(40) | מזהה הלקוח (מפתח זר) |
| IsOrdered | Boolean | האם החיפוש הסתיים בהזמנה או לא |
| **F\_EMPLOYMENT**  טבלת בסיס הנתונים המשני, מציגה מידע שנתי על נתוני התעסוקה במדינות ארה"ב | State (D\_STATES) | Varchar(20) | מזהה המדינה (מפתח זר) |
| Year | Int | מזהה השנה |
| EmployablePopulation | Int | מספר האזרחים במדינה הכשירים לעבודה |
| Employed | Int | מספר המועסקים |
| UnEmployed | Int | מספר המובטלים |
| **F\_SUMMARY**  טבלת עובדה סיכומית, מסכמת את הזמנות החברה יחד עם נתוני התעסוקה לפי מדינה ושנה | State (D\_STATES) | Varchar(20) | מזהה המדינה (מפתח זר) |
| Year | Int | מזהה השנה |
| EmployablePopulation | Int | מספר האזרחים במדינה הכשירים לעבודה |
| Employed | Int | מספר המועסקים |
| UnEmployed | Int | מספר המובטלים |
| Number Of Orders | Int | מספר ההזמנות הכולל  שדה שמחושב ע"י COUNT |
| Total Amount | Money | סכום הקניה הכולל  שדה מחושב ע"י SUM(AMOUNT) |

* לא נדרשו שינויים מיוחדים

# (20%) אפיון תהליכי ה ETL

**Diagram

Description automatically generatedמיפוי מקורות ליעדים**

**הסדר הנכון של הרצת השלבים:**

1. **Data Mirroring (MRR) – שכפול נתוני המקור אל אזור השיקוף**

מטרת שלב זה הינה שכפול מקור הנתונים לאזור השיקוף על מנת לשמור על תקינות נתוני המקור. שלב מקדים לשלב זה הוא יצירת בסיסי נתונים המכילים את נתוני המקור.

1. **Dimension Staging (STG-DIMS) – הכנת נתוני המימד**

מטרת שלב זה הינהביצוע טרנספורמציות לטבלאות המימד על מנת להעביר את הנתונים ממבנה טבלאי לסכמת הכוכב. שלב זה חשוב ביותר על מנת ליישם את סכמת הכוכב.

1. **Dimension Warehousing (DW- DIMS) – עדכון טבלאות המימד ב DW**

מטרת שלב זה הינה העברת הטבלאות החדשות משלב הStaging אל מחסן הנתונים. שלב זה הכרחי ליצירת שדות Valid From, Valid Until אשר נמצאים במימדים משתנים לאט מסוג 2.

1. **Fact Staging (STG-FACT) – הכנת נתוני העובדה**

מטרת שלב זה הינה ביצוע טרנספורמציות על טבלאות העובדה. שלב זה מתבצע באמצעות טבלאות השיקוף מהשלב הראשון ובאמצעות טבלאות המימד מהשלב השלישי.

1. **Referential Integrity (RI) – בדיקת יושרת הנתונים**

מטרת שלב זה הינה בדיקת תקינות הנתונים בטבלאות העובדה שיצרנו בשלב הקודם אל מול הנתונים בטבלאות המימד שהכנסנו למחסן הנתונים בשלב השלישי. שלב זה חשוב לאכיפת יושרת המקורות.

1. **Fact Warehousing (DW-FACT) – עדכון טבלאות העובדה ב DW**

מטרת שלב זה הינה העברת הטבלאות החדשות משלב ה Stagingאל מחסן הנתונים לאחר התיקונים שבוצעו בשלב הקודם. שלב זה הינו השלב הסופי לתהליך ה ETL.

* **התלויות הקיימות בין שלבי התהליך**

כל שלב בתהליך תלוי בשלבים הקודמים לו ולכן יש חשיבות לסדר התהליכים. תחילה, חייב להתקיים השלב הראשון על מנת לא להשתמש בשלבים הבאים במקור הנתונים המקורי. בשלב השני נשתמש בנתונים מהשלב הראשון, בשלב השלישי נשתמש בטבלאות שנוצרו בשלב השני. בשל תלויות מפתחות זרים, השלב הרביעי צריך להתבצע רק לאחר טעינת טבלאות המימד (שלב 3). בשלב החמישי, בדיקת היושרה תתבצע על הטבלאות מהשלבים הקודמים (שלבים 3 ו-4). רק לאחר סיום שלב זה נוכל לבצע את השלב השישי והאחרון בו נעביר את טבלאות העובדה למחסן הנתונים.

* **השלבים שצפויים לקחת זמן רב במיוחד**

1. שלב השיקוף (MRR) – שלב זה ייקח זמן רב מאחר והוא מתבצע על כלל הנתונים הנמצאים במקורות הנתונים.
2. שלב הכנת נתוני העובדה (FACT-STG) – שלב זה צפוי לקחת זמן רב כיוון שתהליך יצירת טבלאות העובדה כולל תחשיבים ופעולות JOIN רבות בין טבלאות שונות.
3. שלב בדיקת יושרת הנתונים (RI) - השלב בו נבדוק את יושרת הנתונים צפוי לקחת זמן רב מאחר והתהליך יכלול מעבר על כלל המפתחות הזרים בנתוני טבלאות העובדה בהשוואה לטבלאות המימד.

* **השלבים שעתידים להיכשל**

1. שלב העתקת הנתונים (MRR) - כחלק מתהליך זה עשויות להופיע שגיאות עקב כמות גדולה של נתונים. על מנת למנוע את שגיאות אלה נבצע העברה בשלבים על פי היררכיה.
2. שלב בדיקת יושרת הנתונים (RI) - כחלק מתהליך זה עשויות להופיע שגיאות עקב רשומות שעבורן לא תימצא התאמה של מפתח בין טבלאות העובדה לטבלאות המימד. על מנת להימנע משגיאות אלו יש להעביר את הרשומות הבעייתיות לטבלת עובדה חדשה או ליצור עבורן רשומות פיקטיביות בטבלאות המימד.

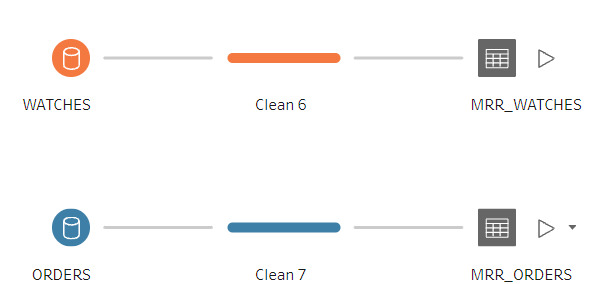
# (40%) מימוש תהליכי ה-ETL – יישומי הבסיס

1. **MRR**

בשלב זה העברנו את הנתונים של כל טבלה מבסיס הנתונים התפעולי לטבלת MRR המתאימה לה.

A picture containing diagram

Description automatically generated



1. **DIMS-STG**

בשלב זה העברנו את הנתונים מטבלאות הMRR לטבלאות ה .STG לחלק מהטבלאות ביצענו JOIN על מנת ליצור את טבלאות אלו עם המפתחות הזרים הרלוונטים.

Diagram

Description automatically generated

עבור טבלת MRR\_STATE\_EMPLOYMENT הסרנו את כל העמודות מלבד עמודת State. ביצענו אגרגציה ע"י GROUP BY לפי State כדי להימנע מליצור רשומות כפולות. לבסוף קיבלנו עמודה של כלל המדינות בארה"ב.

עבור טבלת STG\_CUSTOMERS ביצענו JOIN בין הטבלאות MRR\_CUSTOMERS לבין STG\_STATES על מנת להוסיף את אילוץ המפתח הזר עבור השדה "State".

1. **DIMS-DW**

בשלב זה העברנו את הנתונים מטבלאות הSTG לטבלאות הDIM המתאימות, וניצור את טבלאות המימד בהם נשתמש בשלבים הבאים.

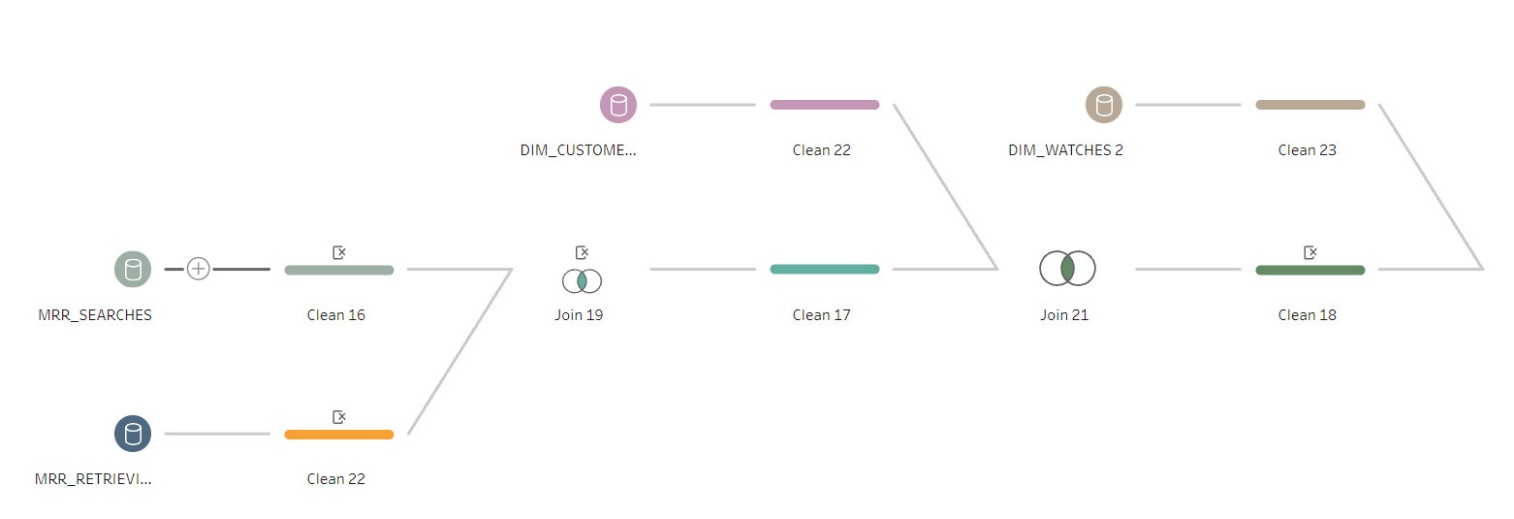
Diagram

Description automatically generated

בנוסף, בשלב זה ניצור את טבלאות DIM\_DATESו DIM\_TIMES באמצעות הקוד הרלוונטי בSQL.

1. **FACT-STG**

בטבלת STG\_SEARCHES התבססנו על טבלת MRR\_SEARCHES וחיברנו באמצעות JOIN את הטבלאות הנדרשות לפי המפתחות הזרים הרלוונטיים תוך הסרת שדות שאינם רלוונטים לטבלה זו:



Chart, funnel chart

Description automatically generated

Diagram

Description automatically generated  
בטבלת STG\_ORDERS\_ITEMS התבבסנו על טבלת MRR\_ORDERS וחיברנו באמצעות JOIN את הטבלאות הנדרשות לפי המפתחות הזרים הרלוונטיים תוך הסרת שדות שאינם רלוונטים לטבלה זו. בנוסף, על מנת לבצע את החישוב עבור השדה ShipPrice לסוג פריט בהזמנה, חישבנו את כמות היחידות שהוזמנו עבור כל הזמנה:

1. **FACT-DW**

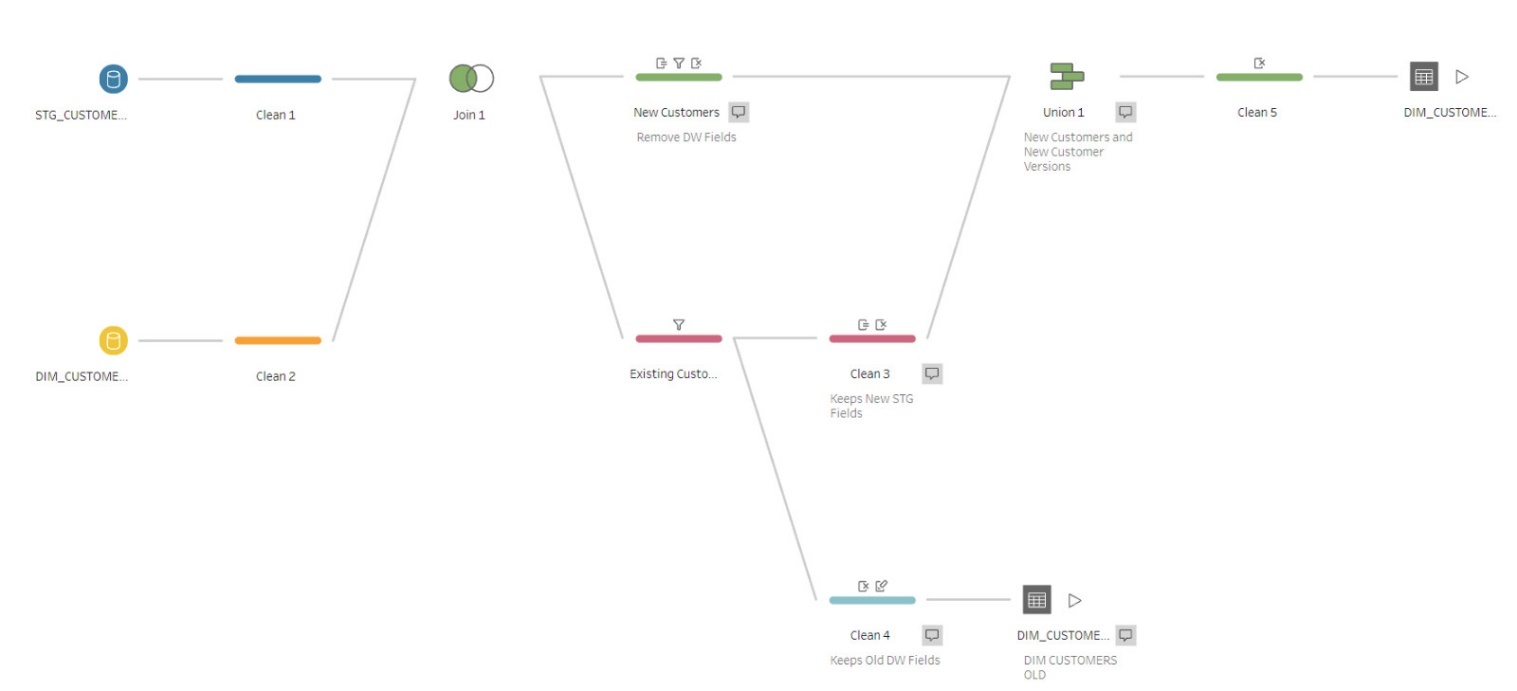
Diagram

Description automatically generated**כעת נעביר את כל טבלאות העובדה המסודרות לטבלאות ה**DW

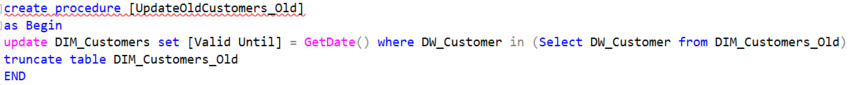
בשלב זה הוספנו לטבלאות העובדה מפתח עזר רלוונטי בעזרת identity(1,1) כפי שלמדנו בSQL. בנוסף, השלמנו את טבלאות הסיכום רק בשלב זה מאחר ואנו צריכים את הנתונים מה-DW.

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

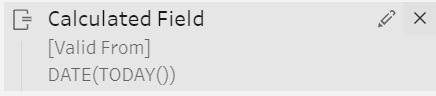
**א. טבלת עובדה מסוג מימד משתנה לאט מסוג 2:**

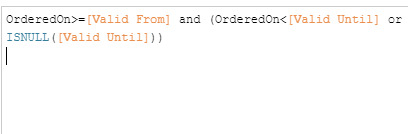


בטבלתDIM\_CUSTOMERS\_OLD ביצענו את הפרוצדורה הבאה לאחר עדכון הטבלה. הפרוצדורה:

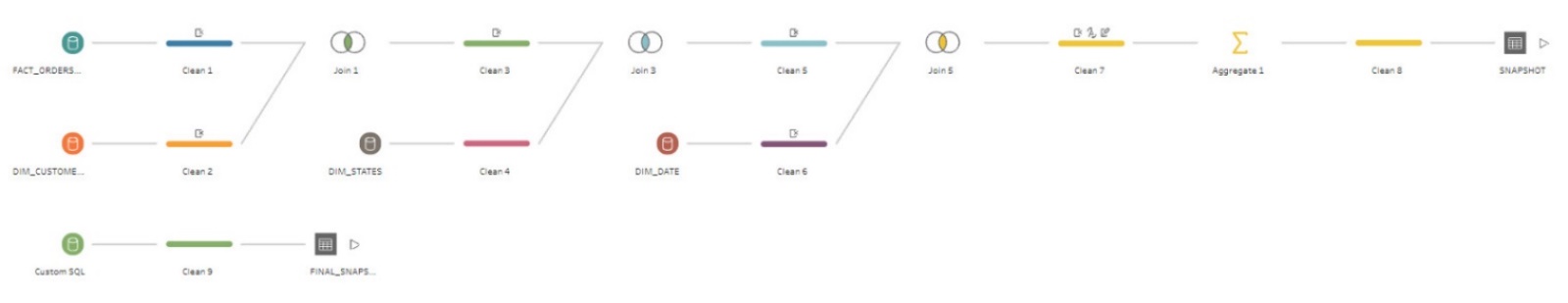


לאחר עדכון הנתונים בעבור לקוח מסוים נרצה לעדכן את שדה ה Valid From ברשומה החדשה לקבל את התאריך הנוכחי.



במידה והיינו עושים את הETL לפי מימד סוג 2, יש להוסיף את הפילטור הבא:  


**ב. טבלת עובדה מסוג Snapshot:**

כל רשומה בטבלה תציג מדינה בתאריך מסוים ואת הערכים המצטברים הבאים: מספר ההזמנות וסכום ההכנסות הכולל. הערכים מצטברים מתחילת השנה עד התאריך של אותה רשומה.

ב FLOW הראשון ביצענו JOIN בין DIM\_ORDERS\_ITEMS לבין DIM\_CUSTOMERSואז ל DIM\_STATES על מנת להוסיף את המפתחות הזרים: DW\_Customers, States. בנוסף ביצענו JOIN בין

DIM\_ORDERS\_ITEMS לבין DIM\_DATES על מנת לחלץ את השנה עבור תאריכי ביצוע ההזמנות

ב FLOW השני ביצענו שאילתה ב Custom SQL:

Graphical user interface, text, application

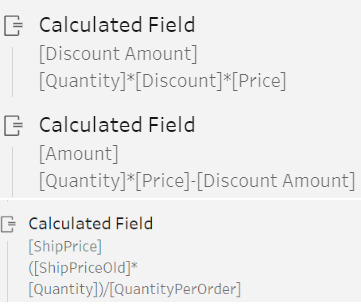
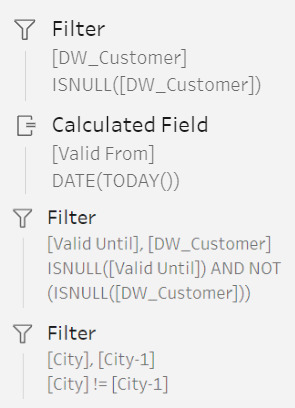
Description automatically generated

בשאילתה זו נבצע צבירה של מספר ההזמנות וסכום ההזמנות הכולל. הצבירה תתבצע עבור כל שנה באמצעות PARTITION BY. סידרנו את הרשומות לכל ”part ”לפי מדינה ושנה.

**נספח 1 - שינויים של פילטר או שדות מחושבים:**

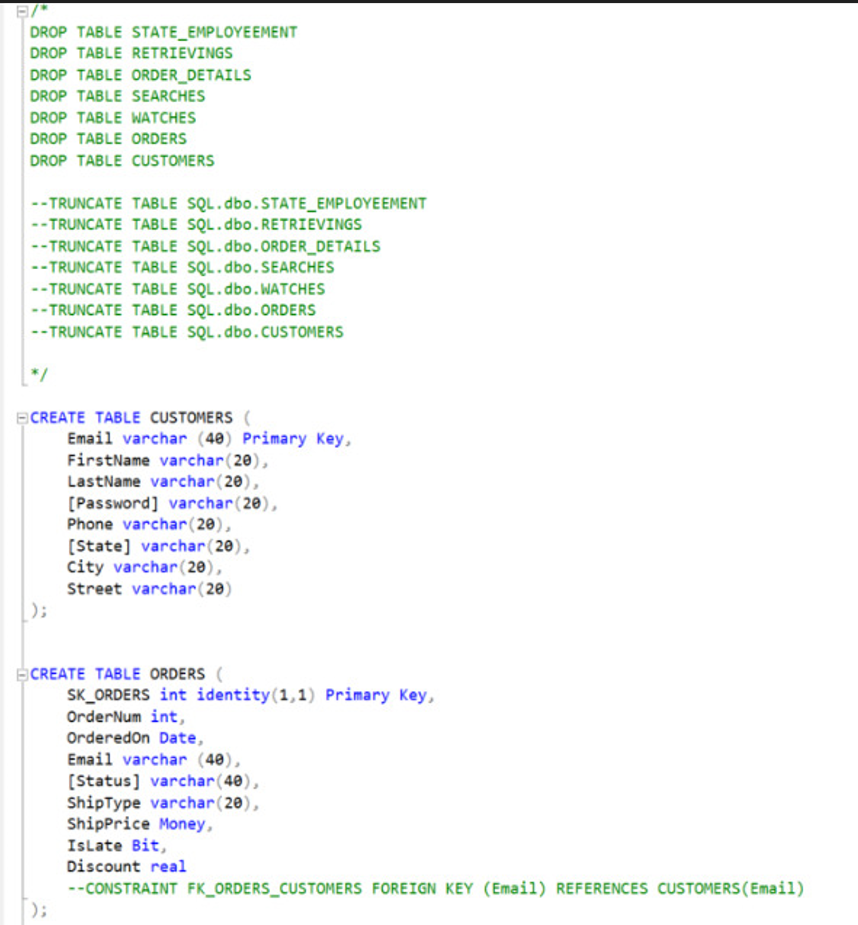
**שדות מחושבים STG-FACT:**

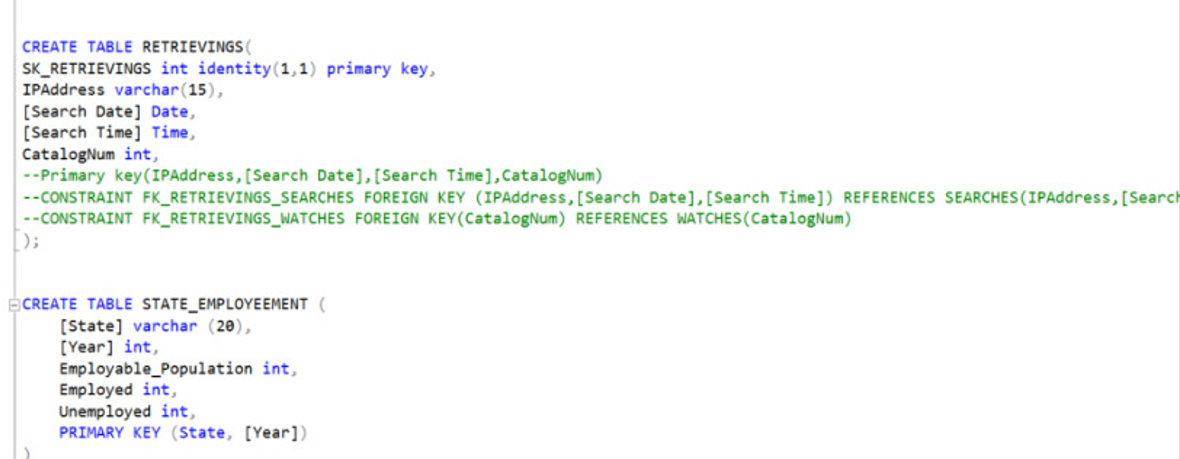
**מימד משתנה לאט מסוג 2:**



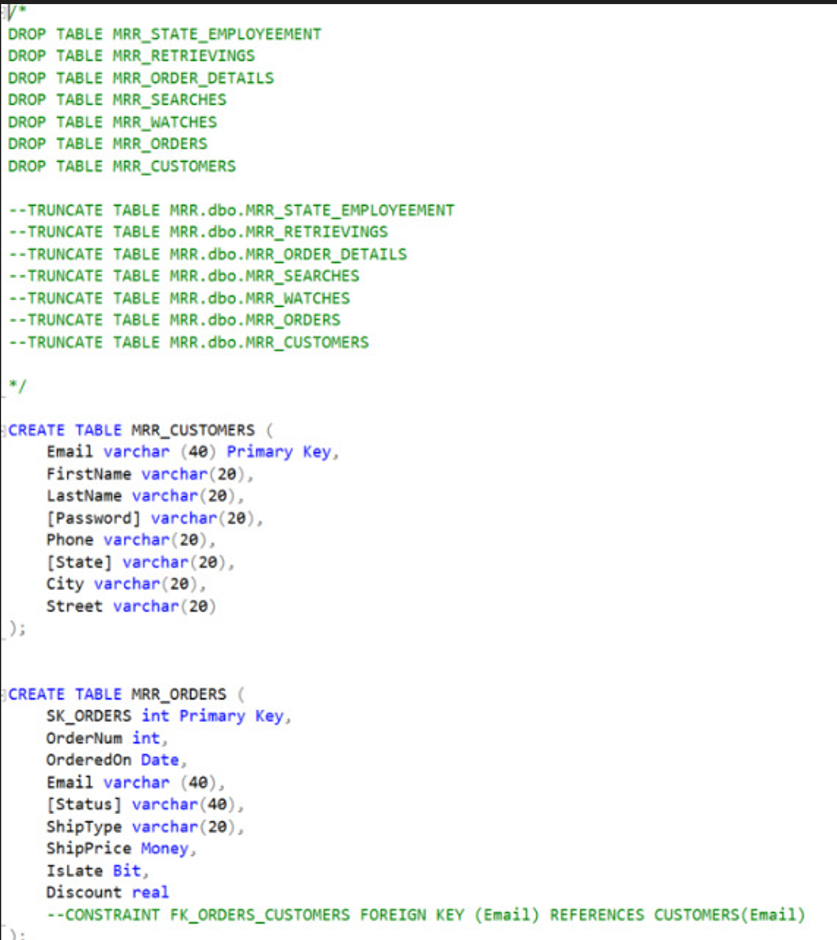
**נספח 2 - קוד ה-SQL היוצר את הטבלאות במחסן הנתונים**

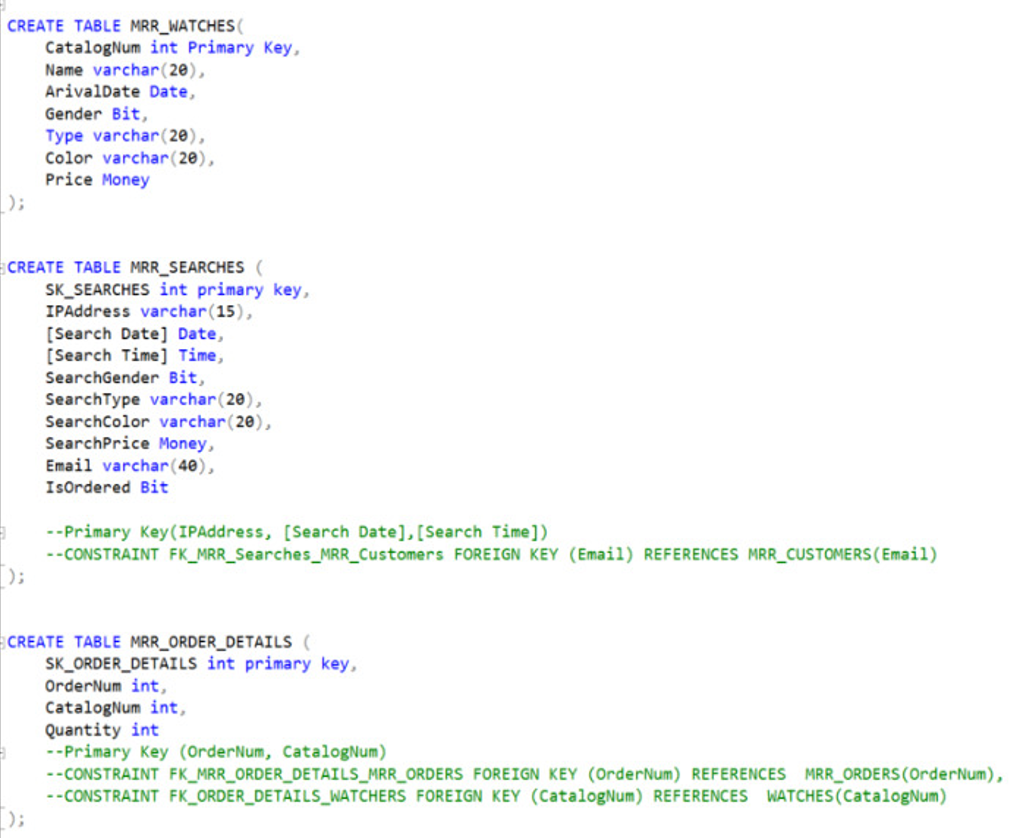
**SQL:**

****

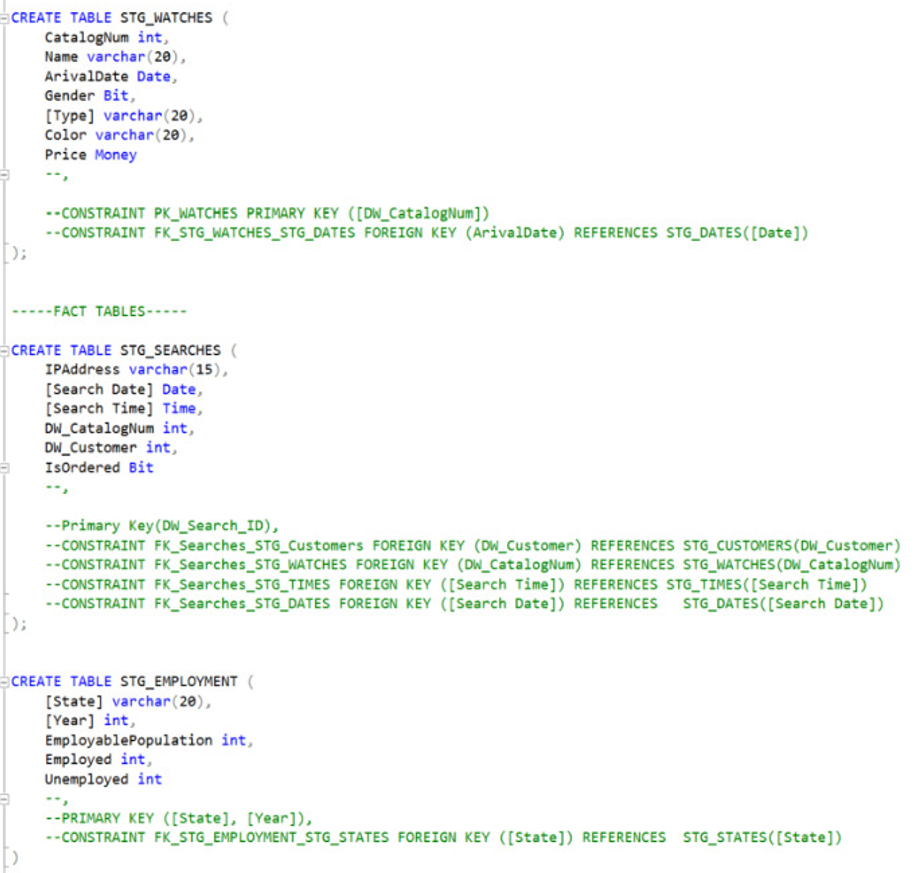
**Graphical user interface, text, application

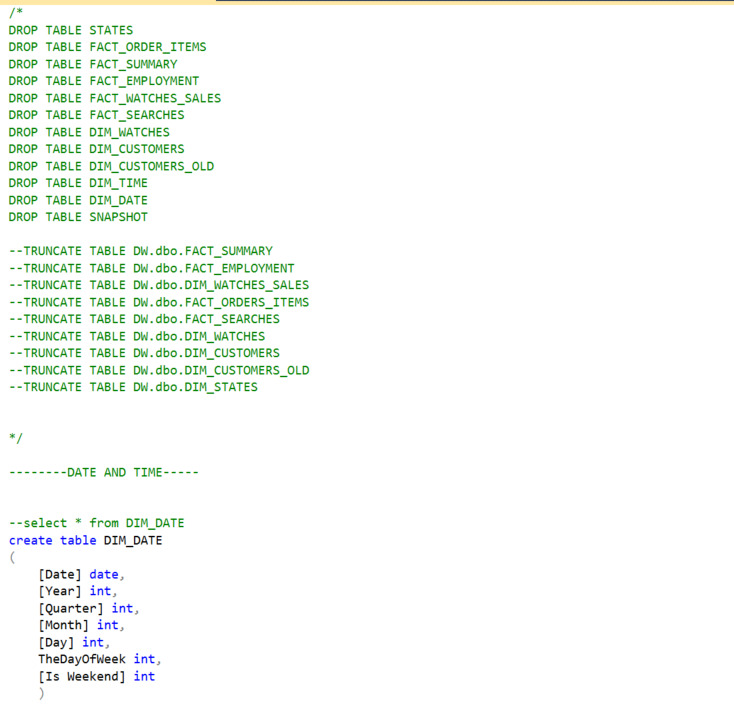
Description automatically generated**

**MRR:**

****

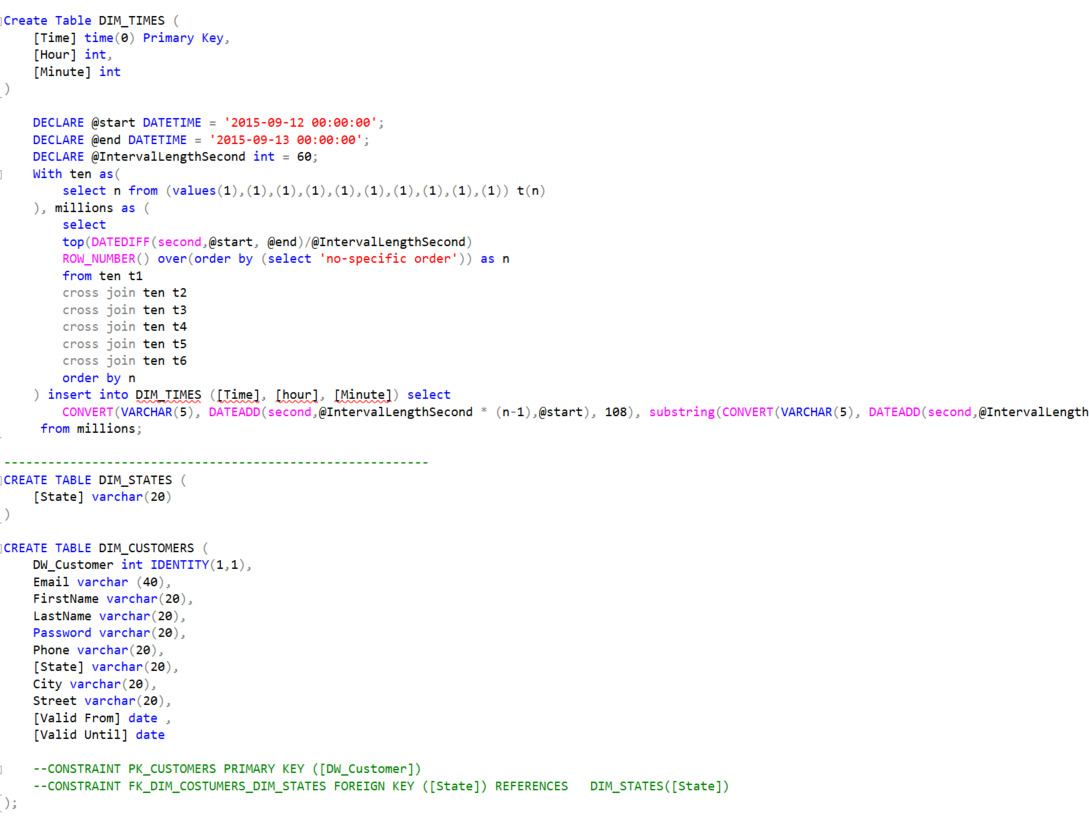
**STG:**

****

**DW:**

Text

Description automatically generated







 Graphical user interface, text, application, email

Description automatically generated