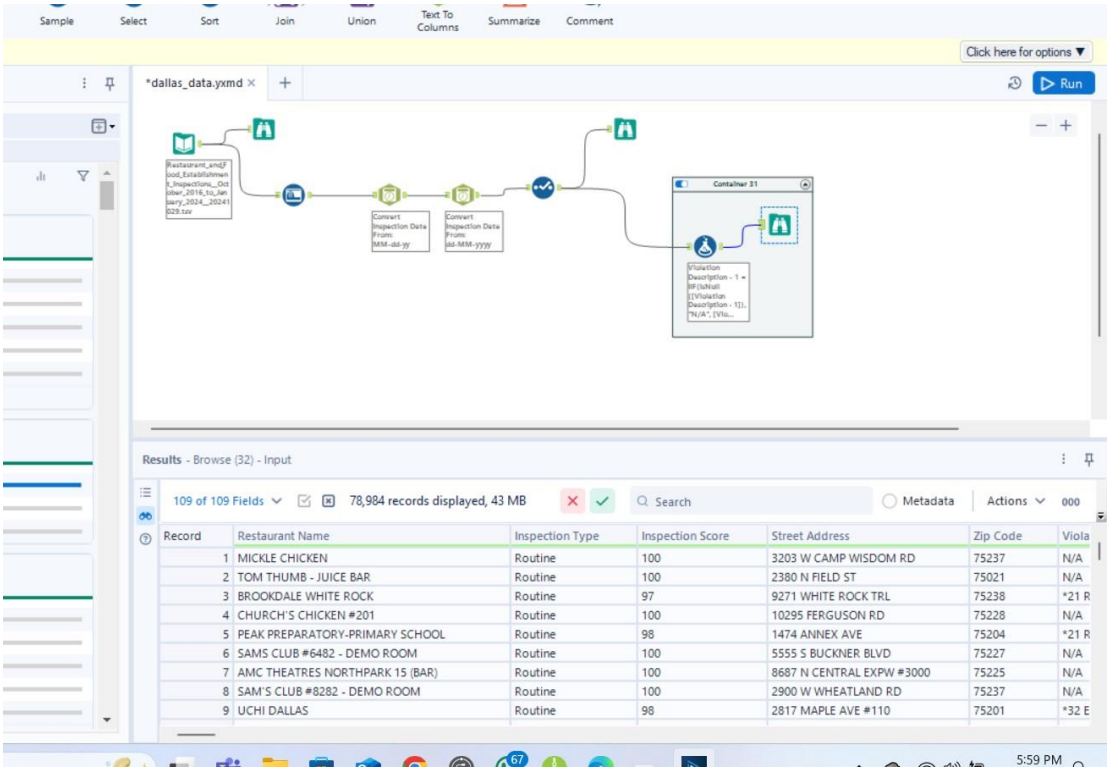
Midterm team 1 Project

By
Dhrumil Patel
Mittul Sharma
Vijeth Reddy

Part 1

- Data Profiling

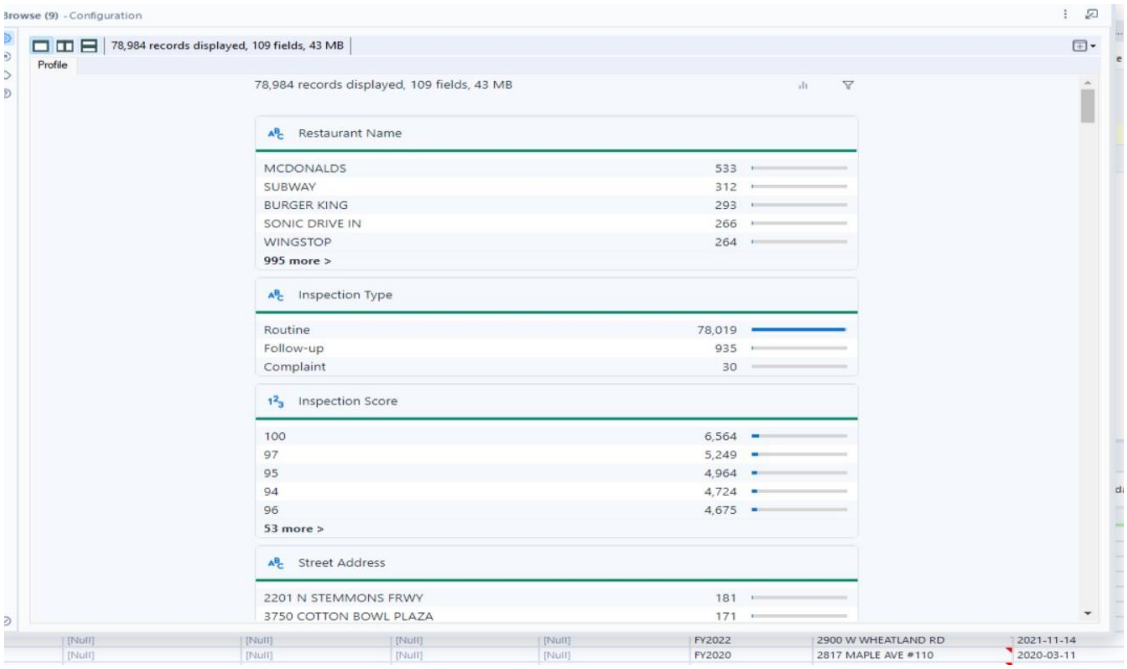


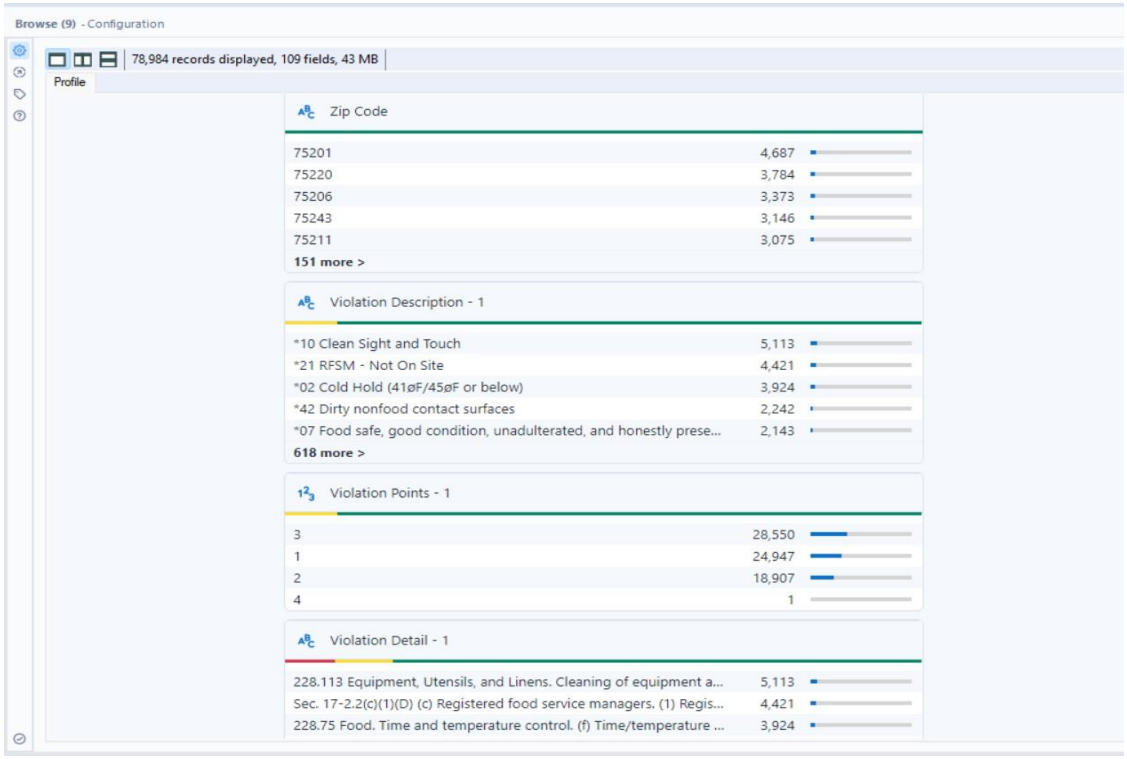
Record	Restaurant Name	Inspection Type	Inspection Score	Street Address	Zip Code	Viola
1	MICKLE CHICKEN	Routine	100	3203 W CAMP WISDOM RD	75237	N/A
2	TOM THUMB - JUICE BAR	Routine	100	2380 N FIELD ST	75021	N/A
3	BROOKDALE WHITE ROCK	Routine	97	9271 WHITE ROCK TRL	75238	*21 R
4	CHURCH'S CHICKEN #201	Routine	100	10295 FERGUSON RD	75228	N/A
5	PEAK PREPARATORY-PRIMARY SCHOOL	Routine	98	1474 ANNEX AVE	75204	*21 R
6	SAMS CLUB #6482 - DEMO ROOM	Routine	100	5555 S BUCKNER BLVD	75227	N/A
7	AMC THEATRES NORTHPARK 15 (BAR)	Routine	100	8687 N CENTRAL EXPW #3000	75225	N/A
8	SAM'S CLUB #8282 - DEMO ROOM	Routine	100	2900 W WHEATLAND RD	75237	N/A
9	UCHI DALLAS	Routine	98	2817 MAPLE AVE #110	75201	*32 E

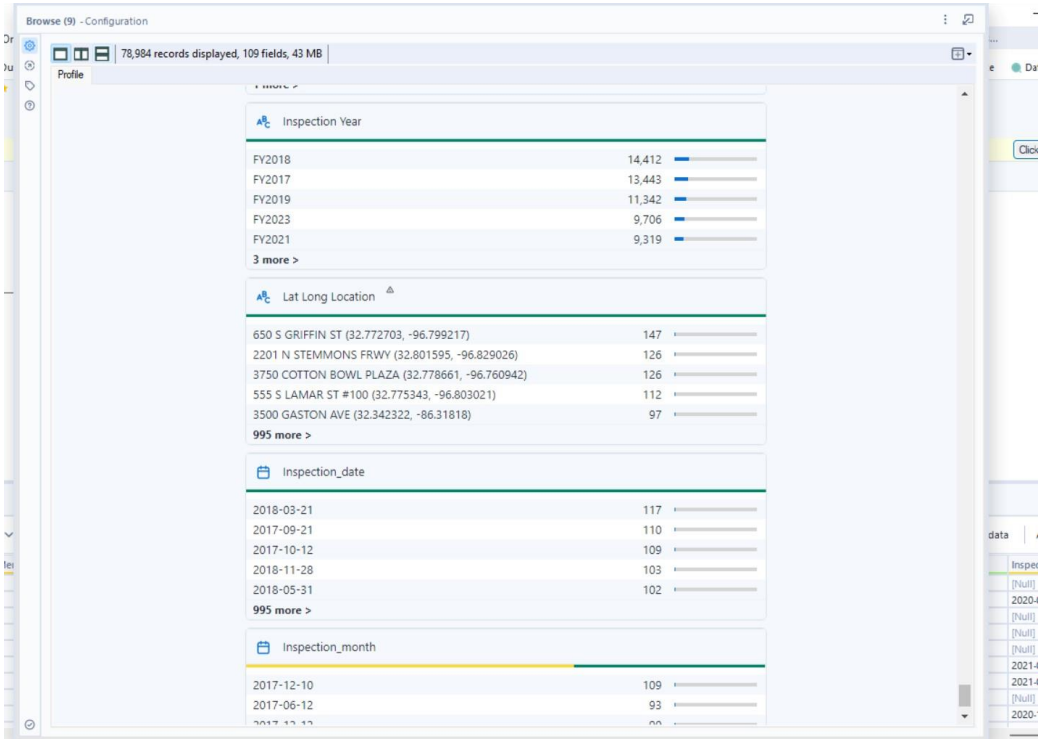
  - Data analysis and key output:
    Columns- there are total 114 columns in Dallas table out of which restaurant name, inspection type, inspection score, street address, zip code, street, inspection date, inspection month, inspection year, location do not have any null values.

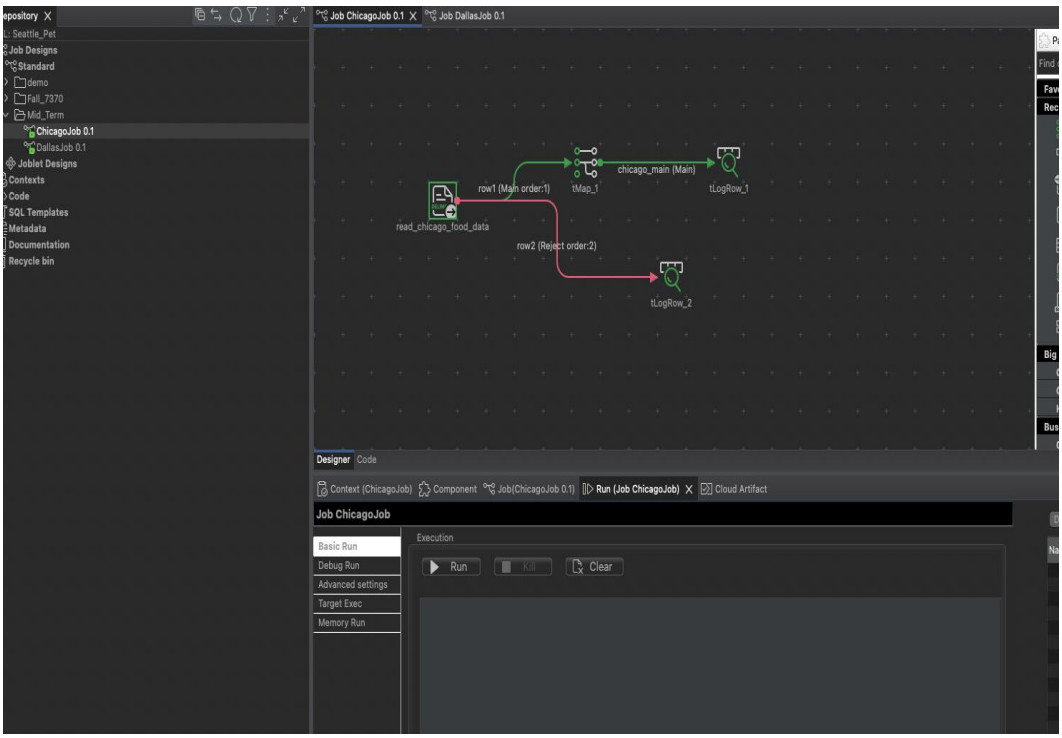
- There is a total of 78984 records in the table.
- Apart from this there is one more data inconsistency such as for violation description, violation details, violation memo, violation points are columns in De-normalize form such as violation description 1, violation description 2, violation description 3 so here is there is violation then it stores in one of column and other columns store null values.

- There are changes of data type of data in inspection date, inspection month, created date in date, zip code, inspection score, violation points in integer and other columns are in string.

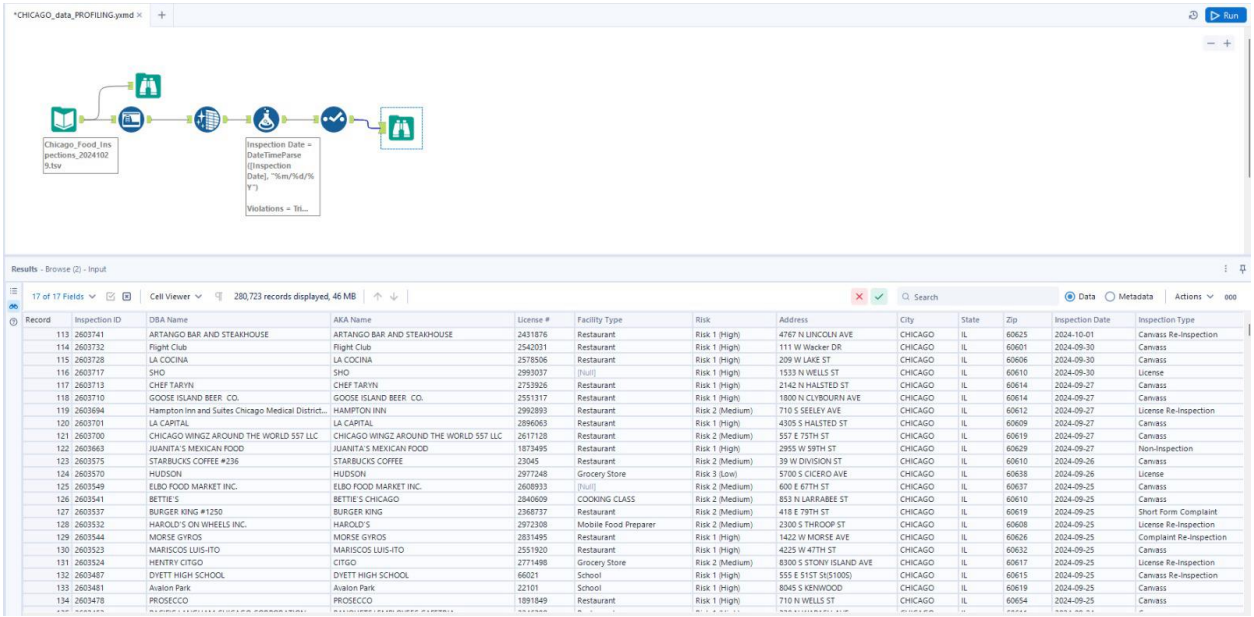



Browse (9) - Configuration
78,984 records displayed, 109 fields, 43 MB
Profile
Zip Code
75201 4,687
75220 3,784
75206 3,373
75243 3,146
75211 3,075
151 more >
Violation Description - 1
*10 Clean Sight and Touch 5,113
*21 RFSM - Not On Site 4,421
*02 Cold Hold (41øF/45øF or below) 3,924
*42 Dirty nonfood contact surfaces 2,242
*07 Food safe, good condition, unadulterated, and honestly prese... 2,143
618 more >
Violation Points - 1
3 28,550
1 24,947
2 18,907
4 1
Violation Detail - 1
228.113 Equipment, Utensils, and Linens. Cleaning of equipment a... 5,113
Sec. 17-2.2(c)(1)(D) (c) Registered food service managers. (1) Regis... 4,421
228.75 Food. Time and temperature control. (f) Time/temperature ... 3,924


Browse (9) - Configuration
78,984 records displayed, 109 fields, 43 MB
Profile
Inspection Year
FY2018 14,412
FY2017 13,443
FY2019 11,342
FY2023 9,706
FY2021 9,319
3 more >
Lat Long Location
650 S GRIFFIN ST (32.772703, -96.799217) 147
2201 N STEMMONS FRWY (32.801595, -96.829026) 126
3750 COTTON BOWL PLAZA (32.778661, -96.760942) 126
555 S LAMAR ST #100 (32.775343, -96.803021) 112
3500 GASTON AVE (32.342322, -86.31818) 97
995 more >
Inspection_date
2018-03-21 117
2017-09-21 110
2017-10-12 109
2018-11-28 103
2018-05-31 102
995 more >
Inspection_month
2017-12-10 109
2017-06-12 93



- Compare to Dallas data Chicago has less data inconsistency
- In Chicago data there is not any denormalized data in this data set there is detailed data is created but when it comes to combine both data some of fields are eliminated such as inspection type, license number
- And here we have used similar data type conversion as per Chicago dataset

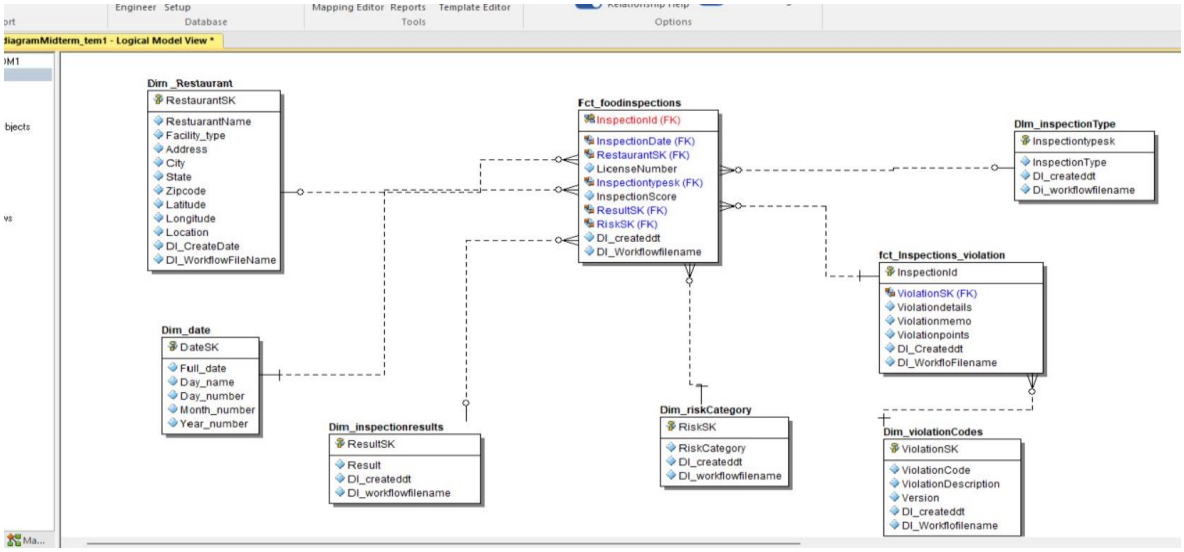


Record	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type
113	2603741	ARTANGO BAR AND STEAKHOUSE	ARTANGO BAR AND STEAKHOUSE	2431876	Restaurant	Risk 1 (High)	4767 N LINCOLN AVE	CHICAGO	IL	60625	2024-10-01	Canvass Re-Inspection
114	2603732	Flight Club	Flight Club	2542031	Restaurant	Risk 1 (High)	111 W Wacker DR	CHICAGO	IL	60601	2024-09-30	Canvass
115	2603728	LA COCINA	LA COCINA	2578506	Restaurant	Risk 1 (High)	209 W LAKE ST	CHICAGO	IL	60606	2024-09-30	Canvass
116	2603717	SHO	SHO	2993037	[Null]	Risk 1 (High)	1533 N WELLS ST	CHICAGO	IL	60610	2024-09-30	License
117	2603713	CHEF TARYN	CHEF TARYN	2753926	Restaurant	Risk 1 (High)	2142 N HALSTED ST	CHICAGO	IL	60614	2024-09-27	Canvass
118	2603710	GOOSE ISLAND BEER CO.	GOOSE ISLAND BEER CO.	2551317	Restaurant	Risk 1 (High)	1800 N CLYBOURN AVE	CHICAGO	IL	60614	2024-09-27	Canvass
119	2603694	Hampton Inn and Suites Chicago Medical District...	HAMPTON INN	2992893	Restaurant	Risk 2 (Medium)	710 S SEELEY AVE	CHICAGO	IL	60612	2024-09-27	License Re-Inspection
120	2603701	LA CAPITAL	LA CAPITAL	2896063	Restaurant	Risk 1 (High)	4305 S HALSTED ST	CHICAGO	IL	60609	2024-09-27	Canvass
121	2603700	CHICAGO WINGZ AROUND THE WORLD 557 LLC	CHICAGO WINGZ AROUND THE WORLD 557 LLC	2617128	Restaurant	Risk 2 (Medium)	557 E 75TH ST	CHICAGO	IL	60619	2024-09-27	Canvass
122	2603663	JUANITA'S MEXICAN FOOD	JUANITA'S MEXICAN FOOD	1873495	Restaurant	Risk 1 (High)	2955 W 59TH ST	CHICAGO	IL	60629	2024-09-27	Non-Inspection
123	2603575	STARBUCKS COFFEE #236	STARBUCKS COFFEE	23045	Restaurant	Risk 2 (Medium)	39 W DIVISION ST	CHICAGO	IL	60610	2024-09-26	Canvass
124	2603570	HUDSON	HUDSON	2977248	Grocery Store	Risk 3 (Low)	5700 S CICERO AVE	CHICAGO	IL	60638	2024-09-26	License
125	2603549	ELBO FOOD MARKET INC.	ELBO FOOD MARKET INC.	2608933	[Null]	Risk 2 (Medium)	600 E 67TH ST	CHICAGO	IL	60637	2024-09-25	Canvass
126	2603541	BETTIE'S	BETTIE'S CHICAGO	2840609	COOKING CLASS	Risk 2 (Medium)	853 N LARRABEE ST	CHICAGO	IL	60610	2024-09-25	Canvass
127	2603537	BURGER KING #1250	BURGER KING	2368737	Restaurant	Risk 2 (Medium)	418 E 79TH ST	CHICAGO	IL	60619	2024-09-25	Short Form Complaint
128	2603532	HAROLD'S ON WHEELS INC.	HAROLD'S	2972308	Mobile Food Preparer	Risk 2 (Medium)	2300 S THROOP ST	CHICAGO	IL	60608	2024-09-25	License Re-Inspection
129	2603544	MORSE GYROS	MORSE GYROS	2831495	Restaurant	Risk 1 (High)	1422 W MORSE AVE	CHICAGO	IL	60626	2024-09-25	Complaint Re-Inspection
130	2603523	MARISCOS LUIS-ITO	MARISCOS LUIS-ITO	2551920	Restaurant	Risk 1 (High)	4225 W 47TH ST	CHICAGO	IL	60632	2024-09-25	Canvass
131	2603524	HENTRY CITGO	CITGO	2771498	Grocery Store	Risk 2 (Medium)	8300 S STONY ISLAND AVE	CHICAGO	IL	60617	2024-09-25	License Re-Inspection
132	2603487	DYETT HIGH SCHOOL	DYETT HIGH SCHOOL	66021	School	Risk 1 (High)	555 E 51ST ST(51005)	CHICAGO	IL	60615	2024-09-25	Canvass Re-Inspection
133	2603481	Avalon Park	Avalon Park	22101	School	Risk 1 (High)	8045 S KENWOOD	CHICAGO	IL	60619	2024-09-25	Canvass
134	2603478	PROSECCO	PROSECCO	1891849	Restaurant	Risk 1 (High)	710 N WELLS ST	CHICAGO	IL	60654	2024-09-25	Canvass

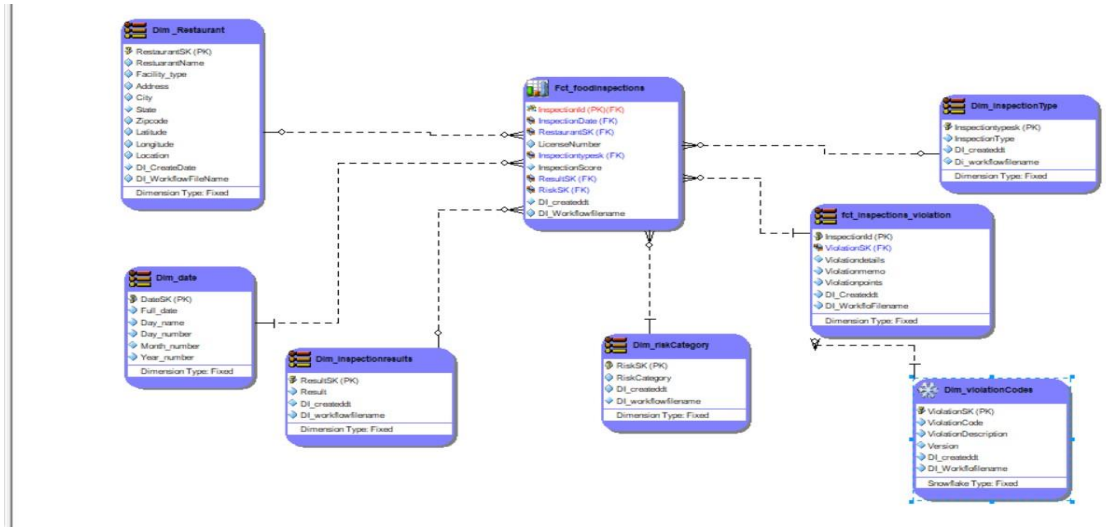
Part 2

- ER Model and DDL Scripts

1. Logical Model



2. Physical Model



3. DDL Script

```sql
-- Drop table Dim_InspectionResults;
CREATE TABLE Dim_InspectionResults (
    ResultsSK INT AUTOINCREMENT(1,1) PRIMARY KEY,
    Results VARCHAR(255),
    DI_CreateDate DATETIME,
    DI_WorkflowFileName STRING
);
-- Drop table Dim_InspectionType ;
CREATE TABLE Dim_InspectionType (
    InspectionTypeSK INT AUTOINCREMENT(1,1) PRIMARY KEY,
    InspectionType VARCHAR(255),
    DI_CreateDate DATETIME,
    DI_WorkflowFileName STRING
);
-- Drop table Dim_RiskCategory ;
CREATE TABLE Dim_RiskCategory (
    RiskSK INT AUTOINCREMENT(1,1) PRIMARY KEY,
    Risk VARCHAR(255),
    DI_CreateDate DATETIME,
    DI_WorkflowFileName STRING
);
-- Drop table Dim_Restaurants ;
CREATE TABLE Dim_Restaurants (
    RestaurantSK INT AUTOINCREMENT(1,1) PRIMARY KEY,
    Restaurant_Name VARCHAR(255),
    Facility_Type VARCHAR(255),
    Address VARCHAR(255),
    City VARCHAR(50),
    State CHAR(10),
    Zip INT,
    Latitude CHAR(18),
    Longitude CHAR(18),
    Location CHAR(40),
    DI_CreateDate DATETIME,
    DI_WorkflowFileName STRING
);

-- Drop table Dim_ViolationCodes ;
CREATE TABLE Dim_ViolationCodes (
    ViolationSK INT AUTOINCREMENT(1,1) PRIMARY KEY,
```

ults Messages

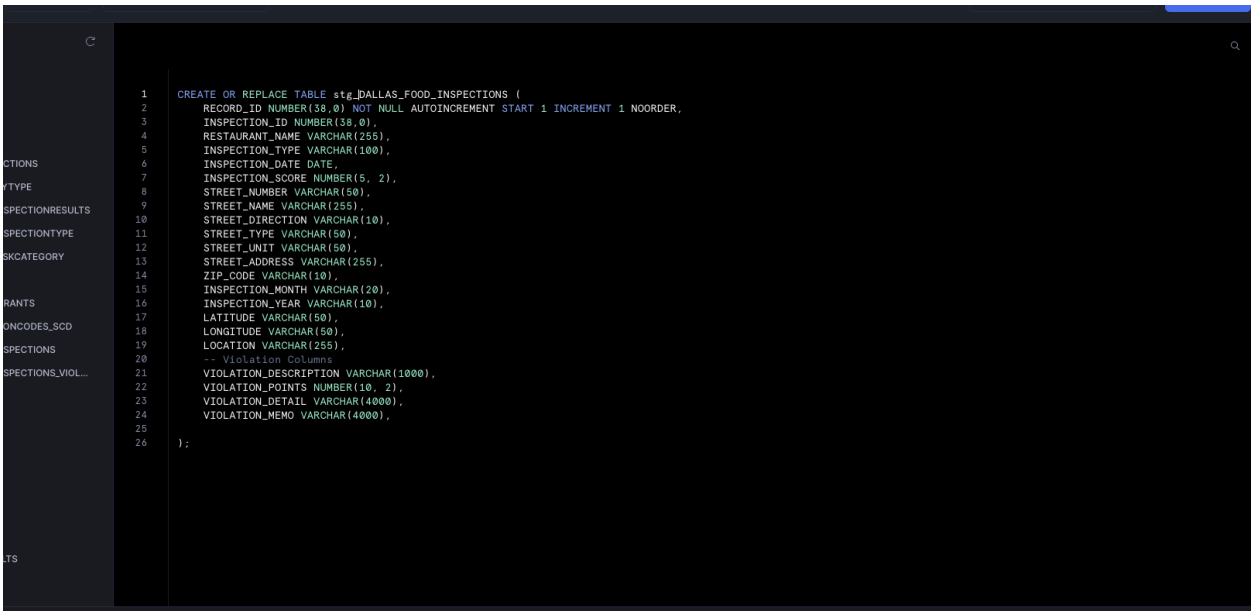
Restaurant_Name	Inspection_Type	Inspection_Date	Inspection_Score	Street_Number	Street_Name	Street_Direction	Street_Type	Street_Unit	Street_Address	Zip_Code	Violation_Description_1
WENDY'S #9780	Routine	2018-09-13	80	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	*31 Handwashing lavatory –
WENDY'S #9780	Routine	2021-04-06	96	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	*32 Maintain in Good Repair
WENDY'S #9780	Routine	2020-02-24	100	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	NULL
WENDY'S #97800	Routine	2021-12-22	92	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75215-4514	*19 Plumbing System Construc
WENDY'S #9780	Routine	2016-12-13	92	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	*31 Individual, disposable

```sql
-- Drop table Dim_ViolationCodes ;
CREATE TABLE Dim_ViolationCodes (
    ViolationSK INT AUTOINCREMENT(1,1) PRIMARY KEY,
    ViolationCode INT,
    ViolationDescription VARCHAR(1000),
     DI_CreateDate DATETIME,
    DI_WorkflowFileName STRING

);
Drop table FCT_Inspections_Violations ;

CREATE TABLE FCT_Inspections_Violations (
    InspectionID INT,                               -- Foreign key linking to FCT_FoodInspections
    ViolationSK INT,                                -- Foreign key referencing Dim_ViolationCodes(ViolationSK)
    ViolationDetail VARCHAR(8000) NULL,                  -- Detailed description of the violation
    ViolationMemo VARCHAR(8000) NULL,                    -- Additional comments or memo
    ViolationPoints Varchar(100) NULL,                        -- Points related to the violation (set as INT, adjust if necessary)
    DI_CreateDate DATETIME,                         -- Creation timestamp for ETL audit
    DI_WorkflowFileName STRING,                     -- Workflow file name for ETL audit
    PRIMARY KEY (InspectionID, ViolationSK),        -- Composite primary key
    FOREIGN KEY (ViolationSK) REFERENCES Dim_ViolationCodes(ViolationSK)  -- Foreign key constraint
);


Select * from FCT_FoodInspections ;

CREATE TABLE FCT_FoodInspections (
    FCTInspectionID INT AUTOINCREMENT(1,1) PRIMARY KEY,
    InspectionID  INT,
    InspectionDate DATE NULL,
    DateSK INT,
    RestaurantSK INT NULL,
    InspectionTypeSK INT NULL,
    ResultsSK INT NULL,
    RiskSK INT NULL,
    Inspection_Score INT NULL,
    DI_CreateDate DATETIME,
    DI_WorkflowFileName STRING
```

sults Messages

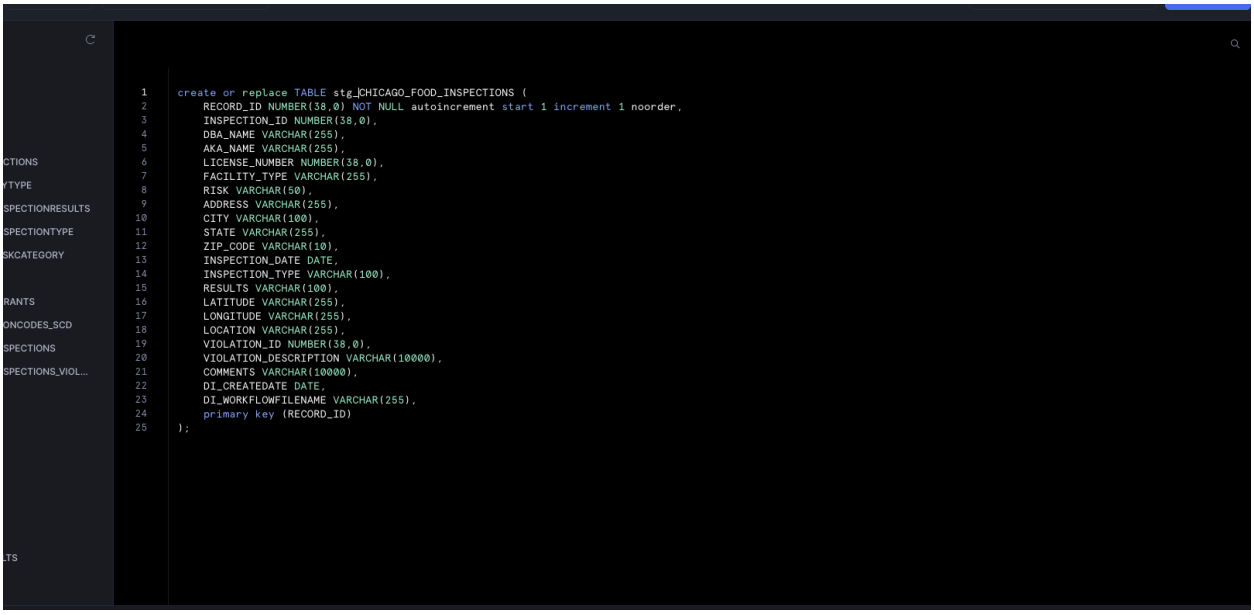
Restaurant_Name	Inspection_Type	Inspection_Date	Inspection_Score	Street_Number	Street_Name	Street_Direction	Street_Type	Street_Unit	Street_Address	Zip_Code	Violation_Description_1
WENDY'S #9780	Routine	2018-09-13	80	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	*31 Handwashing lavatory –
WENDY'S #9780	Routine	2021-04-06	96	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	*32 Maintain in Good Repai
WENDY'S #9780	Routine	2020-02-24	100	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	NULL
WENDY'S #97800	Routine	2021-12-22	92	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75215-4514	*19 Plumbing System Constr
WENDY'S #9780	Routine	2016-12-13	92	1507	KIEST	E	BLVD	NULL	1507 E KIEST BLVD	75216	*31 Individual, disposable

4. Dallas Stage Table



```
CREATE OR REPLACE TABLE stg_DALLAS_FOOD_INSPECTIONS (
    RECORD_ID NUMBER(38,0) NOT NULL AUTOINCREMENT START 1 INCREMENT 1 NOORDER,
    INSPECTION_ID NUMBER(38,0),
    RESTAURANT_NAME VARCHAR(255),
    INSPECTION_TYPE VARCHAR(100),
    INSPECTION_DATE DATE,
    INSPECTION_SCORE NUMBER(5, 2),
    STREET_NUMBER VARCHAR(50),
    STREET_NAME VARCHAR(255),
    STREET_DIRECTION VARCHAR(10),
    STREET_TYPE VARCHAR(50),
    STREET_UNIT VARCHAR(50),
    STREET_ADDRESS VARCHAR(255),
    ZIP_CODE VARCHAR(10),
    INSPECTION_MONTH VARCHAR(20),
    INSPECTION_YEAR VARCHAR(10),
    LATITUDE VARCHAR(50),
    LONGITUDE VARCHAR(50),
    LOCATION VARCHAR(255),
    -- Violation Columns
    VIOLATION_DESCRIPTION VARCHAR(1000),
    VIOLATION_POINTS NUMBER(10, 2),
    VIOLATION_DETAIL VARCHAR(4000),
    VIOLATION_MEMO VARCHAR(4000),

);
```

5. Chicago Stage Table



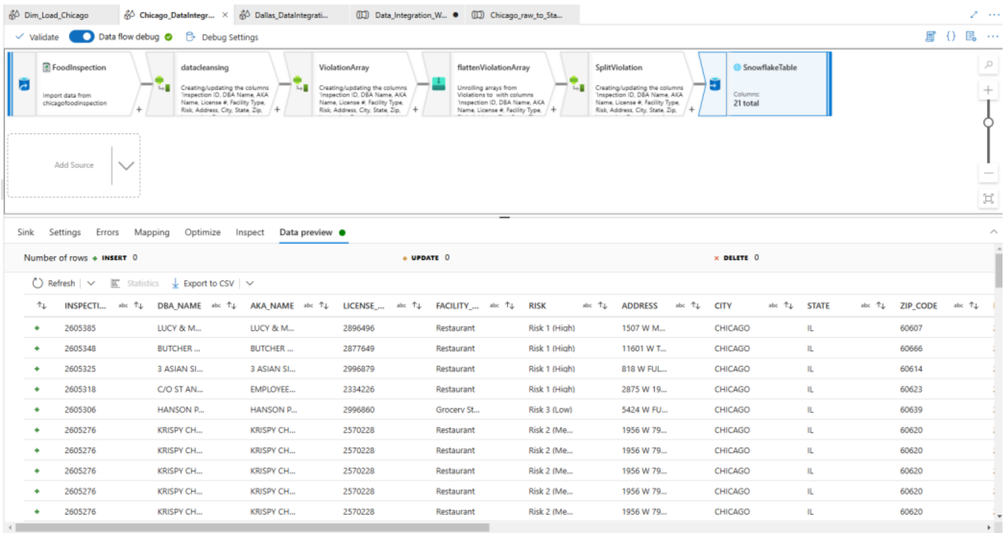
```
create or replace TABLE stg_CHICAGO_FOOD_INSPECTIONS (
    RECORD_ID NUMBER(38,0) NOT NULL autoincrement start 1 increment 1 noorder,
    INSPECTION_ID NUMBER(38,0),
    DBA_NAME VARCHAR(255),
    AKA_NAME VARCHAR(255),
    LICENSE_NUMBER NUMBER(38,0),
    FACILITY_TYPE VARCHAR(255),
    RISK VARCHAR(50),
    ADDRESS VARCHAR(255),
    CITY VARCHAR(100),
    STATE VARCHAR(255),
    ZIP_CODE VARCHAR(10),
    INSPECTION_DATE DATE,
    INSPECTION_TYPE VARCHAR(100),
    RESULTS VARCHAR(100),
    LATITUDE VARCHAR(255),
    LONGITUDE VARCHAR(255),
    LOCATION VARCHAR(255),
    VIOLATION_ID NUMBER(38,0),
    VIOLATION_DESCRIPTION VARCHAR(10000),
    COMMENTS VARCHAR(10000),
    DI_CREATEDATE DATE,
    DI_WORKFLOWFILENAME VARCHAR(255),
    primary key (RECORD_ID)
);
```

Part 3

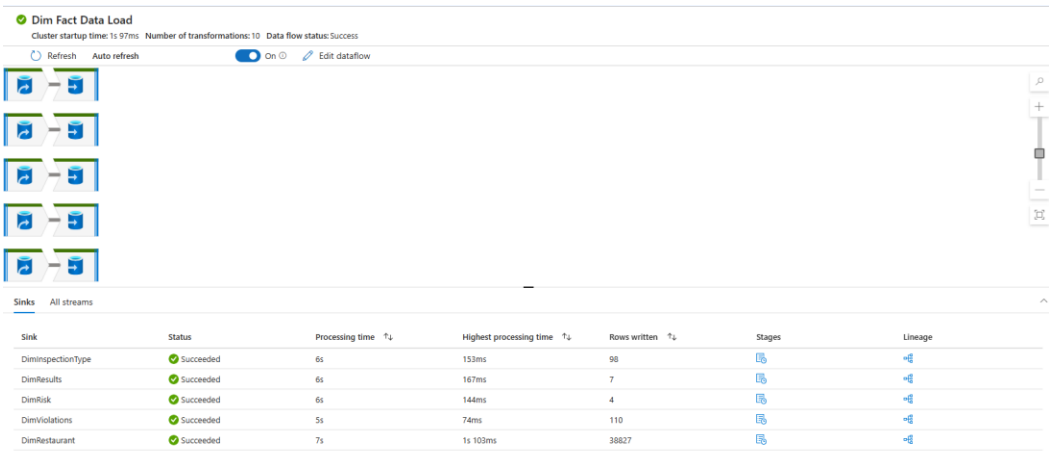
Integration and snowflake

- Chicago stage table in snowflake

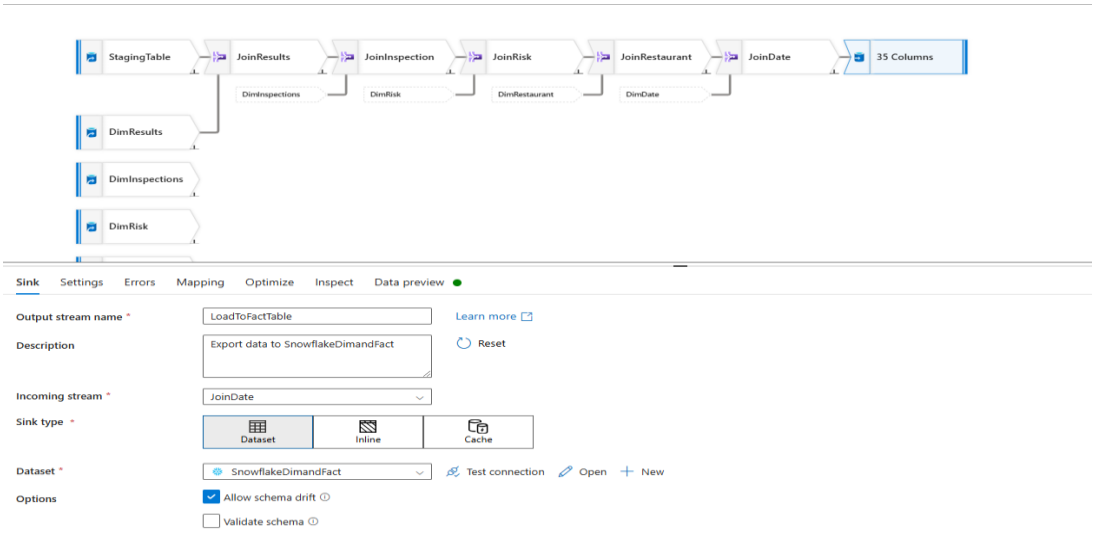
1. Data Integration Workflow to load datasets into Snowflake as staging table



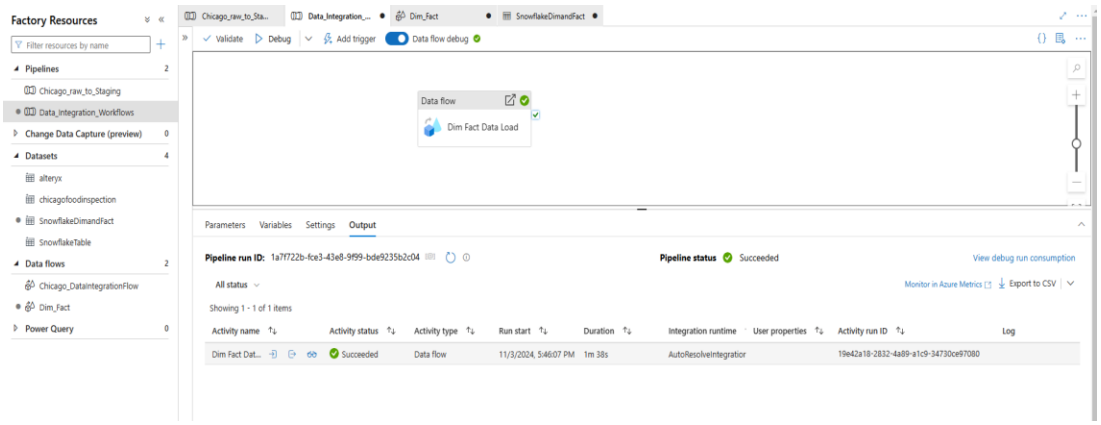
2. Load Staging data into Dimensional tables



3. Pipeline to load data into fact table from dimensional table



4. Snapshot of Staging to Dimension tables



To merge the stg_DALLAS_FOOD_INSPECTIONS and stg_CHICAGO_FOOD_INSPECTIONS datasets, here's the plan:

1. **Align Column Names**: Map similar columns, like RESTAURANT_NAME (Dallas) to DBA_NAME (Chicago) and consolidate address components into a single ADDRESS field.
2. **Include Unique Columns**: Add Chicago-specific fields (e.g., LICENSE_NUMBER, FACILITY_TYPE, RISK, RESULTS) to the merged schema, with null values allowed for Dallas entries.
3. **Standardize Violation Details**: Combine violation columns across datasets, such as VIOLATION_DESCRIPTION and COMMENTS.
4. **Unified Schema**: Create a schema that captures fields from both datasets, allowing flexibility in columns exclusive to either Dallas or Chicago.
5. Merge similar columns: Map VIOLATION_DESCRIPTION from both datasets into a unified VIOLATION_DESCRIPTION field.

6. Include VIOLATION_POINTS (Dallas-only) and VIOLATION_ID (Chicago-only) in the merged schema, allowing nulls where they don't apply.
7. Consolidate VIOLATION_DETAIL (Dallas) and COMMENTS (Chicago) into a single VIOLATION_DETAIL field, storing specific notes or remarks related to the violation.
8. Include VIOLATION_MEMO (Dallas-only) for additional detail where available.
9. Design a unified violation section in the merged schema to capture all aspects of violation data, ensuring any city-specific fields have null values for records from the other city.