Sri Lanka Institute of Information Technology

IT3021 - Data warehousing and Business Intelligence

Year 3 Semester 2

DWBI – Assignment 01

Prepared by: Nishadi Jayarathna.

Step 1: Scenario Description

The **Farm-to-Table Produce Delivery Tracker** dataset represents an organic produce delivery service that connects local farmers to retailers, such as grocery stores and restaurants, to facilitate the sale and delivery of fresh, organic crops like vegetables, fruits, and herbs. The dataset tracks the entire lifecycle of the process, including crop production, order placement, delivery logistics, and payment transactions, over the course of one year **(January 1, 2024, to December 31, 2024)**. This scenario is **novel because it integrates agriculture (crop production and farmer certification), logistics (delivery routes and delays), and business analytics (sales and performance tracking), offering a unique context compared to typical retail or e-commerce datasets.**

Key entities in the dataset include:

- **Farmers:** Details of 100 farmers, including their names, farm names, locations, organic certification status (Certified, Pending, Non-Certified), and farm sizes.
- **Products:** Information on 200 organic produce items (e.g., kale, strawberries), including categories (e.g., Vegetables, Fruits), subcategories (e.g., Leafy Greens, Berries), unit prices, and seasonal availability (e.g., Spring, Year-Round).
- **Retailers:** Details of 200 retailers, including their names, types (e.g., Grocery, Restaurant), and locations.
- **Orders:** 50,000 order transactions, capturing order dates, total amounts, and links to farmers and retailers.
- **Order Details:** 150,000-line items specifying products and quantities within orders.
- **Delivery Routes:** 50,000 delivery records, including route IDs, drivers, delivery status (On-Time, Delayed), and delay reasons.
- **Payments:** 50,000 payment transactions, including amounts and payment methods.
- **Accumulating Transaction Updates:** 50,000 records providing completion timestamps for orders, used to calculate processing times.

The dataset is stored in three formats to simulate real-world heterogeneous systems:

- **Excel:** Static reference data for **farmers, products, and retailers**, resembling a master database. [Farmers.xlsx, Products.xlsx, Retailers.xlsx]
- **TXT: Transactional logs for orders and order details**, mimicking an order management system. [Orders.txt, OrderDetails.txt]
- **CSV:** External system exports for **delivery routes, payments, and transaction updates and supports tracking for Organic Certification Impact on Sales**, representing logistics and payment systems. [DeliveryRoutes.csv, Payments.csv, FarmerUpdates.csv, AccmTxnUpdates.csv - Step 6]

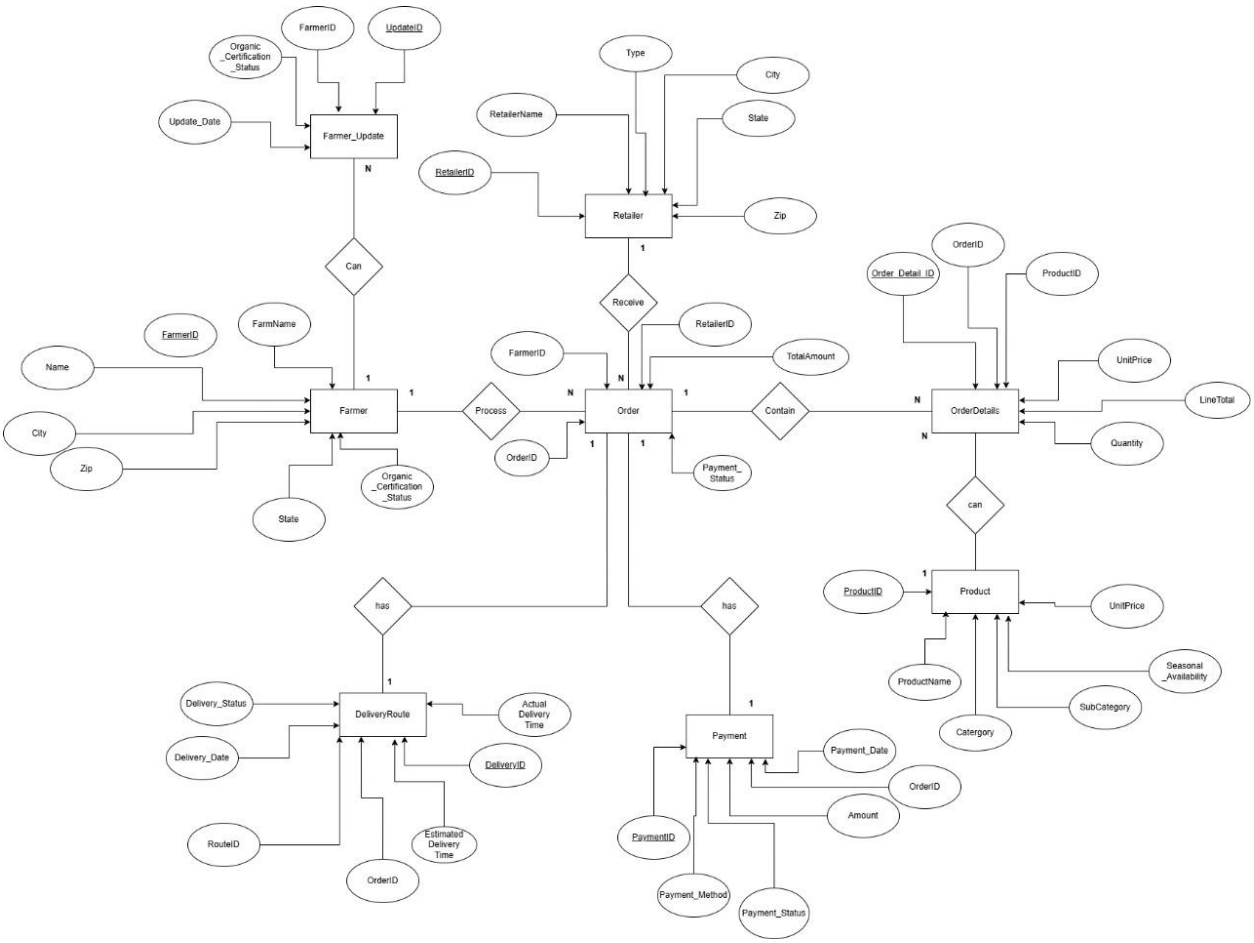
Providing sufficient data to demonstrate data warehousing concepts, including:

- **Dimensions and Hierarchies:** Farmer (State → City → Farmer), Product (Category → SubCategory → Product), Retailer (State → City → Retailer), and Date (Year → Quarter → Month → Day).
- **Slowly Changing Dimension:** Farmer's **"OrganicCertificationStatus,City,FarmName,FarmSize,Name**" for SCD tracking.
  Product's **"Seasonal Availability, Unit Price"** for SCD tracking.
  Delivery Route's **"Actual Delivery Time, Delivery Date, Estimated Delivery"** for SCD tracking.
- **Fact Table:** OrderSalesFact captures transactional sales data, including key measures such as TotalAmount, Quantity, and order Status. It integrates foreign keys to related dimensions like FarmerSK, ProductSK, RetailerSK, DeliverySK, and DateKey for analytical slicing. The table also supports payment analysis through fields like PaymentStatus, PaymentID, and PaymentMethod, enabling insights into transaction trends and payment preferences.
- **ETL Processes:** Extracting from Excel, TXT, and CSV, transforming, and loading into a star schema.

- **Analytics:** Sales by product category, delivery efficiency, and farmer performance, Retailer Sales Performance by Type, Top-Performing Products by Revenue, Regional Sales Distribution Organic Certification Impact on Sales, payment method trend analysis via SSAS cubes and reports.

The dataset's volume (50,000 orders, 150,000 order details) and time span (2024) enable seasonal and regional analysis, making it ideal for building a robust data warehouse and generating actionable business insights, such as optimizing delivery routes or identifying high-demand products.

ER Diagram



Step 2: Preparation of Data Sources

Data Source Organization

The dataset is organized into three folders under **C:\IT3021_DataSources\:**

- **Excel: C:\IT3021_DataSources\Excel** for static reference data (Farmers.xlsx, Products.xlsx, Retailers.xlsx).
- **TXT: C:\IT3021_DataSources\TXT** for transactional logs (Orders.txt, OrderDetails.txt).
- **CSV: C:\IT3021_DataSources\CSV** for external system exports (DeliveryRoutes.csv, Payments.csv, FarmerUpdates.csv, AccmTxnUpdates.csv).

Source Descriptions

1. **Excel Files:**

   - **Farmers.xlsx:**

     - **Information:** Contains details of 100 farmers, including FarmerID (primary key), Name, FarmName, City, State, Zip, OrganicCertificationStatus (Certified, Pending, Non-Certified), and FarmSize (acres).
     - **Format:** Excel (.xlsx), single worksheet, 100 records.
     - **Purpose:** Source for FarmerDim, supports a hierarchy (State → City → Farmer) and Type 1 SCD with OrganicCertificationStatus,City,FarmName,FarmSize,Name.
     - **Key Attributes:** FarmerID, OrganicCertificationStatus, State, City.

   - **Products.xlsx:**

     - **Information:** Contains details of 200 organic produce items, including ProductID (primary key), ProductName, Category (e.g., Vegetables), SubCategory (e.g., Leafy Greens), UnitPrice, and SeasonalAvailability (e.g., Spring, Year-Round).
     - **Format:** Excel (.xlsx), single worksheet, 200 records.

- **Purpose:** Source for ProductDim, supports a hierarchy (Category → SubCategory → Product) and seasonal analysis and Type 1 SCD with Seasonal Availability, Unit Price.
- **Key Attributes:** ProductID, Category, SubCategory, SeasonalAvailability.

- **Retailers.xlsx:**

  - **Information:** Contains details of 200 retailers, including RetailerID (primary key), RetailerName, Type (e.g., Grocery, Restaurant), City, State, and Zip.
  - **Format:** Excel (.xlsx), single worksheet, 200 records.
  - **Purpose:** Source for RetailerDim, supports a hierarchy (State → City → Retailer).
  - **Key Attributes:** RetailerID, Type, State, City.

2. TXT Files:

- **Orders.txt:**

- **Information:** Contains 50,000 order transactions, including OrderID (primary key), FarmerID (foreign key to Farmers.xlsx), RetailerID (foreign key to Retailers.xlsx), OrderDate, TotalAmount, Status, and PaymentStatus.
- **Format:** Tab-delimited text (.txt), 50,000 records.
- **Purpose:** Primary source for OrderSalesFact, providing measures like TotalAmount and accm_txn_create_time (set to OrderDate).
- **Key Attributes:** OrderID, TotalAmount, OrderDate, FarmerID, RetailerID.

- **OrderDetails.txt:**

- **Information:** Contains 150,000 order line items, including OrderDetailID (primary key), OrderID (foreign key to Orders.txt), ProductID (foreign key to Products.xlsx), Quantity, UnitPrice, and LineTotal.
- **Format:** Tab-delimited text (.txt), 150,000 records.
- **Purpose:** Provides detailed measures (e.g., Quantity, LineTotal) for OrderSalesFact, supporting product-level sales analysis.
- **Key Attributes:** OrderID, ProductID, Quantity, LineTotal.

3. CSV Files:

- **DeliveryRoutes.csv:**

- **Information:** Contains 50,000 delivery records, including DeliveryID (primary key), OrderID (foreign key to Orders.txt), RouteID, DriverID, DeliveryDate, EstimatedDeliveryTime, ActualDeliveryTime, DeliveryStatus (On-Time, Delayed), and DelayReason.
- **Format:** Comma-separated (.csv), 50,000 records.
- **Purpose:** Supports delivery performance analytics (e.g., delay trends) and can be joined with OrderSalesFact for logistics insights.
- **Key Attributes:** OrderID, DeliveryStatus, DelayReason.

- **Payments.csv:**

- **Information:** Contains 50,000 payment transactions, including PaymentID (primary key), OrderID (foreign key to Orders.txt), PaymentDate, Amount, PaymentMethod, and PaymentStatus.
- **Format:** Comma-separated (.csv), 50,000 records.
- **Purpose:** Supports financial analytics (e.g., payment trends) and can be joined with OrderSalesFact.
- **Key Attributes:** OrderID, Amount, PaymentMethod.

- **FarmerUpdates.csv**:

- **Information**: Contains records of updates to farmer attributes, specifically changes to OrganicCertificationStatus, including FarmerID (foreign key to **Farmers.xlsx**), OrganicCertificationStatus, UpdateDate, and optional UpdateID (unique identifier for each update).
- **Format**: Comma-separated (.csv), assumed to contain 100+ records (to match or extend the 100 farmers in **Farmers.xlsx** and capture historical changes).
- **Purpose**: Supports historical tracking of farmer attributes for analytical reports, particularly **Organic Certification Impact on Sales**, by enabling Type 1 SCD in **FarmerDim**.
- **Key Attributes**: FarmerID, OrganicCertificationStatus, UpdateDate.

- **AccmTxnUpdates.csv (Step 6):**

- **Information:** Contains 50,000 transaction update records, including txn_id (foreign key matching OrderID in Orders.txt) and accm_txn_complete_time (timestamp of order completion).
- **Format:** Comma-separated (.csv), 50,000 records.
- **Purpose:** Updates OrderSalesFact with accm_txn_complete_time and enables calculation of txn_process_time (hours between accm_txn_create_time and accm_txn_complete_time) for Step 6, supporting insights into order processing efficiency.
- **Key Attributes:** txn_id, accm_txn_complete_time.

Data Relationships

Orders.txt links to:

- "Farmers.xlsx" via "FarmerID".

- "Retailers.xlsx" via "RetailerID".

OrderDetails.txt links to:

- "Orders.txt" via "OrderID".

- "Products.xlsx" via "ProductID".

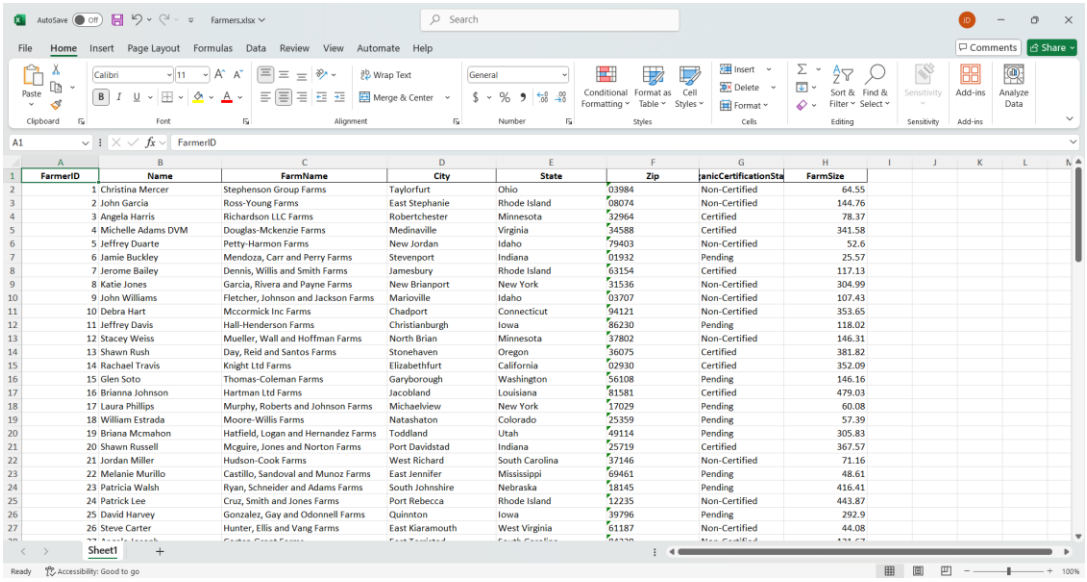
DeliveryRoutes.csv and **Payments.csv** link to "Orders.txt" via "OrderID".

FarmerUpdates.csv links to:

- **Farmers.xlsx** via FarmerID.
- **FarmerDim** via FarmerID (used in ETL to populate Type 1 SCD attributes: OrganicCertificationStatus,City,FarmName,FarmSize,Name).

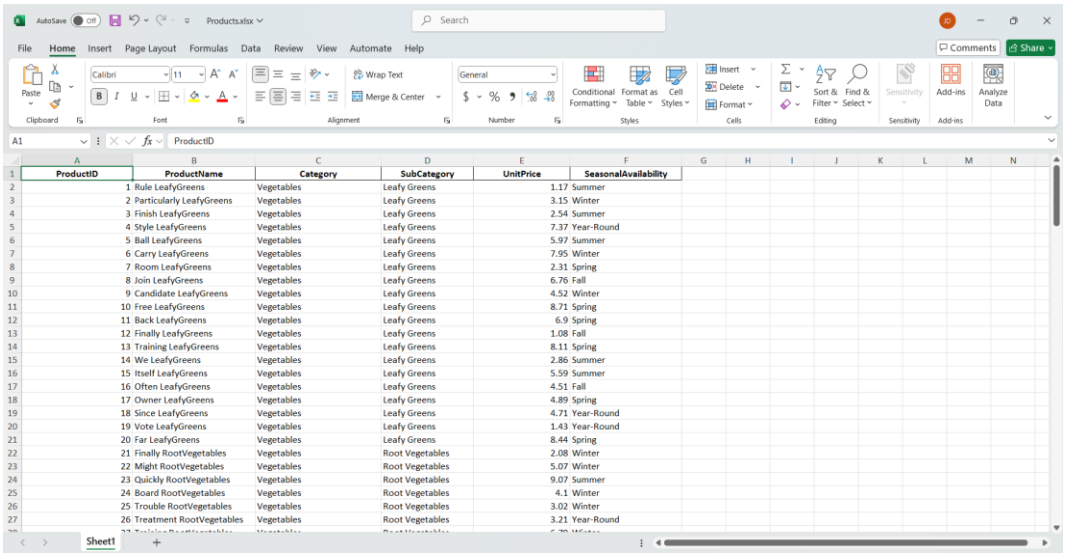
AccmTxnUpdates.csv (Step 6) links to "Orders.txt" and "OrderSalesFact" via "txn_id", which matches "OrderID".

Visual Evidence



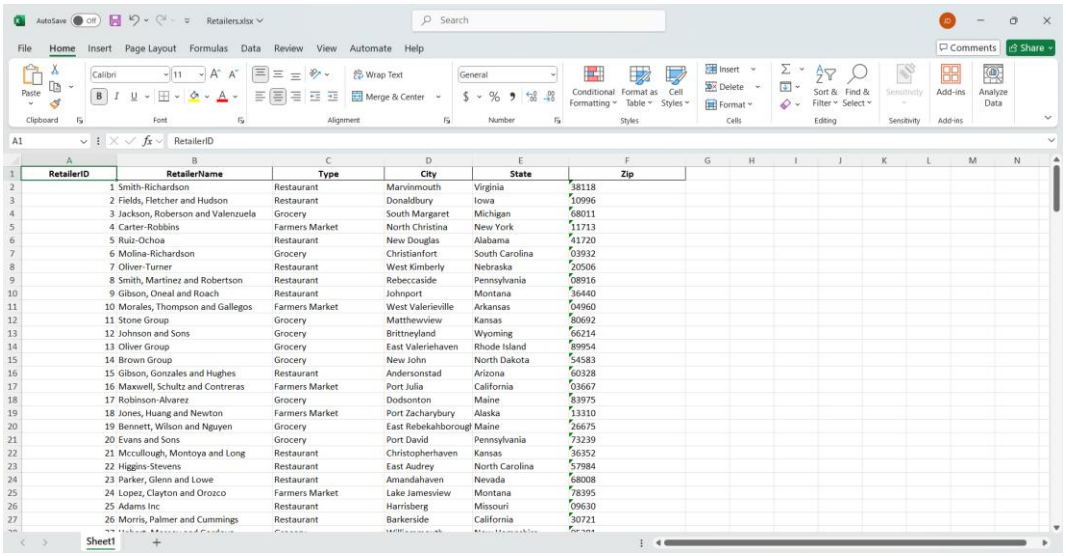
FarmerID	Name	FarmName	City	State	Zip	anicCertificationSta	FarmSize
1	Christina Mercer	Stephenson Group Farms	Taylorfurt	Ohio	03984	Non-Certified	64.55
2	John Garcia	Ross-Young Farms	East Stephanie	Rhode Island	08074	Non-Certified	144.76
3	Angela Harris	Richardson LLC Farms	Robertchester	Minnesota	32964	Certified	78.37
4	Michelle Adams DVM	Douglas-Mckenzie Farms	Medinaville	Virginia	34588	Certified	341.58
5	Jeffrey Duarte	Petty-Harmon Farms	New Jordan	Idaho	79403	Non-Certified	52.6
6	Jamie Buckley	Mendoza, Carr and Perry Farms	Stevenport	Indiana	01932	Pending	25.57
7	Jerome Bailey	Dennis, Willis and Smith Farms	Jamesbury	Rhode Island	63154	Certified	117.13
8	Katie Jones	Garcia, Rivera and Payne Farms	New Brianport	New York	31536	Non-Certified	304.99
9	John Williams	Fletcher, Johnson and Jackson Farms	Marioville	Idaho	03707	Non-Certified	107.43
10	Debra Hart	Mccormick Inc Farms	Chadport	Connecticut	94121	Non-Certified	353.65
11	Jeffrey Davis	Hall-Henderson Farms	Christianburgh	Iowa	86230	Pending	118.02
12	Stacey Weiss	Mueller, Wall and Hoffman Farms	North Brian	Minnesota	37802	Non-Certified	146.31
13	Shawn Rush	Day, Reid and Santos Farms	Stonehaven	Oregon	36075	Certified	381.82
14	Rachael Travis	Knight Ltd Farms	Elizabethfurt	California	02930	Certified	352.09
15	Glen Soto	Thomas-Coleman Farms	Garyborough	Washington	56108	Pending	146.16
16	Brianna Johnson	Hartman Ltd Farms	Jacobland	Louisiana	81581	Certified	479.03
17	Laura Phillips	Murphy, Roberts and Johnson Farms	Michaelview	New York	17029	Pending	60.08
18	William Estrada	Moore-Willis Farms	Natashaton	Colorado	25359	Pending	57.39
19	Briana Mcmahon	Hatfield, Logan and Hernandez Farms	Toddland	Utah	49114	Pending	305.83
20	Shawn Russell	Mcguire, Jones and Norton Farms	Port Davidstad	Indiana	25719	Certified	367.57
21	Jordan Miller	Hudson-Cook Farms	West Richard	South Carolina	37146	Non-Certified	71.16
22	Melanie Murillo	Castillo, Sandoval and Munoz Farms	East Jennifer	Mississippi	69461	Pending	48.61
23	Patricia Walsh	Ryan, Schneider and Adams Farms	South Johnshire	Nebraska	18145	Pending	416.41
24	Patrick Lee	Cruz, Smith and Jones Farms	Port Rebecca	Rhode Island	12235	Non-Certified	443.87
25	David Harvey	Gonzalez, Gay and Odonnell Farms	Quinnton	Iowa	39796	Pending	292.9
26	Steve Carter	Hunter, Ellis and Vang Farms	East Kiaramouth	West Virginia	61187	Non-Certified	44.08

Figure 1: Farmers.xlsx



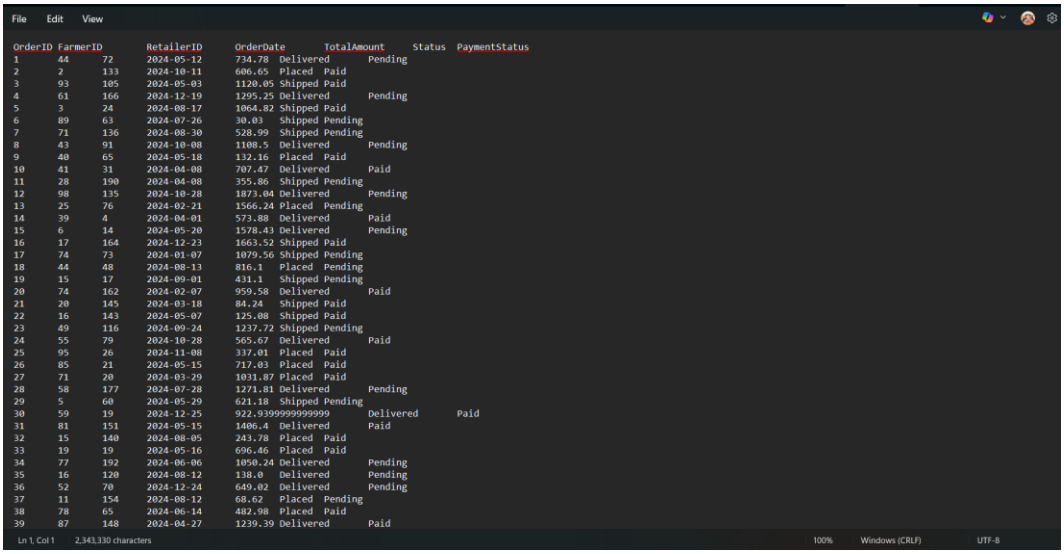
ProductID	ProductName	Category	SubCategory	UnitPrice	SeasonalAvailability
1	Rule LeafyGreens	Vegetables	Leafy Greens	1.17	Summer
2	Particularly LeafyGreens	Vegetables	Leafy Greens	3.15	Winter
3	Finish LeafyGreens	Vegetables	Leafy Greens	2.54	Summer
4	Style LeafyGreens	Vegetables	Leafy Greens	7.37	Year-Round
5	Ball LeafyGreens	Vegetables	Leafy Greens	5.97	Summer
6	Carry LeafyGreens	Vegetables	Leafy Greens	7.95	Winter
7	Room LeafyGreens	Vegetables	Leafy Greens	2.31	Spring
8	Join LeafyGreens	Vegetables	Leafy Greens	6.76	Fall
9	Candidate LeafyGreens	Vegetables	Leafy Greens	4.52	Winter
10	Free LeafyGreens	Vegetables	Leafy Greens	8.71	Spring
11	Back LeafyGreens	Vegetables	Leafy Greens	6.9	Spring
12	Finally LeafyGreens	Vegetables	Leafy Greens	1.08	Fall
13	Training LeafyGreens	Vegetables	Leafy Greens	8.11	Spring
14	We LeafyGreens	Vegetables	Leafy Greens	2.86	Summer
15	Itself LeafyGreens	Vegetables	Leafy Greens	5.59	Summer
16	Often LeafyGreens	Vegetables	Leafy Greens	4.51	Fall
17	Owner LeafyGreens	Vegetables	Leafy Greens	4.89	Spring
18	Since LeafyGreens	Vegetables	Leafy Greens	4.71	Year-Round
19	Vote LeafyGreens	Vegetables	Leafy Greens	1.43	Year-Round
20	Far LeafyGreens	Vegetables	Leafy Greens	8.44	Spring
21	Finally RootVegetables	Vegetables	Root Vegetables	2.08	Winter
22	Might RootVegetables	Vegetables	Root Vegetables	5.07	Winter
23	Quickly RootVegetables	Vegetables	Root Vegetables	9.07	Summer
24	Board RootVegetables	Vegetables	Root Vegetables	4.1	Winter
25	Trouble RootVegetables	Vegetables	Root Vegetables	3.02	Winter
26	Treatment RootVegetables	Vegetables	Root Vegetables	3.21	Year-Round

Figure 2: Products.xlsx



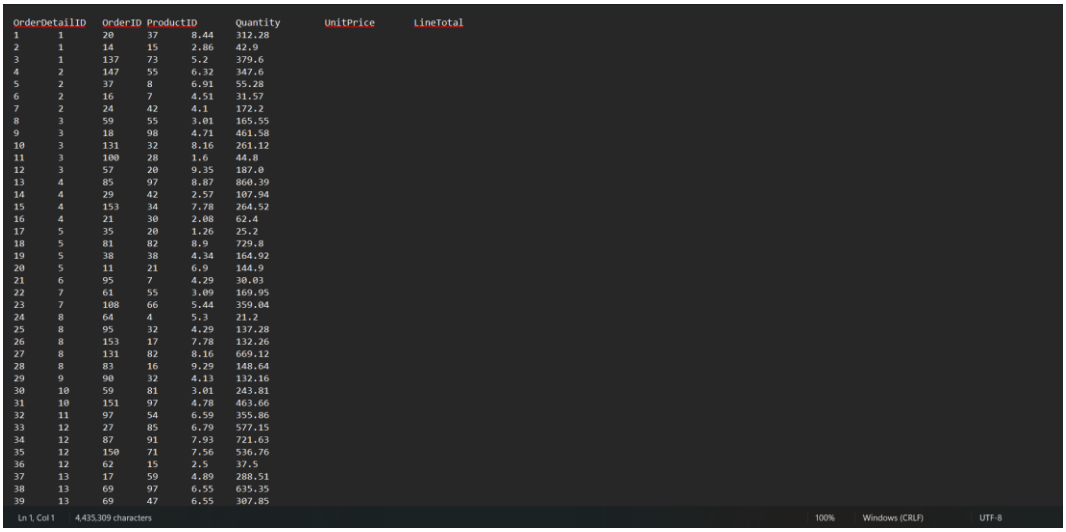
RetailerID	RetailerName	Type	City	State	Zip
1	Smith-Richardson	Restaurant	Marvinmouth	Virginia	38118
2	Fields, Fletcher and Hudson	Restaurant	Donaldbury	Iowa	10996
3	Jackson, Roberson and Valenzuela	Grocery	South Margaret	Michigan	68011
4	Carter-Robbins	Farmers Market	North Christina	New York	11713
5	Ruiz-Ochoa	Restaurant	New Douglas	Alabama	41720
6	Molina-Richardson	Grocery	Christianfort	South Carolina	03932
7	Oliver-Turner	Restaurant	West Kimberly	Nebraska	20506
8	Smith, Martinez and Robertson	Restaurant	Rebeccaside	Pennsylvania	08916
9	Gibson, Oneal and Roach	Restaurant	Johnport	Montana	36440
10	Morales, Thompson and Gallegos	Farmers Market	West Valerieville	Arkansas	04960
11	Stone Group	Grocery	Matthewview	Kansas	80692
12	Johnson and Sons	Grocery	Brittneyland	Wyoming	66214
13	Oliver Group	Grocery	East Valeriehaven	Rhode Island	89954
14	Brown Group	Grocery	New John	North Dakota	54583
15	Gibson, Gonzales and Hughes	Restaurant	Andersonstad	Arizona	60328
16	Maxwell, Schultz and Contreras	Farmers Market	Port Julia	California	03667
17	Robinson-Alvarez	Grocery	Dodsonton	Maine	83975
18	Jones, Huang and Newton	Farmers Market	Port Zacharybury	Alaska	13310
19	Bennett, Wilson and Nguyen	Grocery	East Rebekahborough	Maine	26675
20	Evans and Sons	Grocery	Port David	Pennsylvania	73239
21	Mccullough, Montoya and Long	Restaurant	Christopherhaven	Kansas	36352
22	Higgins-Stevens	Restaurant	East Audrey	North Carolina	57984
23	Parker, Glenn and Lowe	Restaurant	Amandahaven	Nevada	68008
24	Lopez, Clayton and Orozco	Farmers Market	Lake Jamesview	Montana	78395
25	Adams Inc	Restaurant	Harrisberg	Missouri	09630
26	Morris, Palmer and Cummings	Restaurant	Barkerside	California	30721

Figure 3: Retailers.xlsx



OrderID	FarmerID	RetailerID	OrderDate	TotalAmount	Status	PaymentStatus
1	44	72	2024-05-12	734.78	Delivered	Pending
2	2	133	2024-10-11	606.65	Placed	Paid
3	93	105	2024-05-03	1120.05	Shipped	Paid
4	61	166	2024-12-19	1295.25	Delivered	Pending
5	3	24	2024-08-17	1064.82	Shipped	Paid
6	89	63	2024-07-26	30.03	Shipped	Pending
7	71	136	2024-08-30	528.99	Shipped	Pending
8	43	91	2024-10-08	1108.5	Delivered	Pending
9	40	65	2024-05-18	132.16	Placed	Paid
10	41	31	2024-04-08	707.47	Delivered	Paid
11	28	190	2024-04-08	355.86	Shipped	Pending
12	98	135	2024-10-28	1873.04	Delivered	Pending
13	25	76	2024-02-21	1566.24	Placed	Pending
14	39	4	2024-04-01	573.88	Delivered	Paid
15	6	14	2024-05-20	1578.43	Delivered	Pending
16	17	164	2024-12-23	1663.52	Shipped	Paid
17	74	73	2024-01-07	1079.56	Shipped	Pending
18	44	48	2024-08-13	816.1	Placed	Pending
19	15	17	2024-09-01	431.1	Shipped	Pending
20	74	162	2024-02-07	959.58	Delivered	Paid
21	20	145	2024-03-18	84.24	Shipped	Paid
22	16	143	2024-05-07	125.08	Shipped	Paid
23	49	116	2024-09-24	1237.72	Shipped	Pending
24	55	79	2024-10-28	565.67	Delivered	Paid
25	95	26	2024-11-08	337.01	Placed	Paid
26	85	21	2024-05-15	717.03	Placed	Paid
27	71	20	2024-03-29	1031.87	Placed	Paid
28	58	177	2024-07-28	1271.81	Delivered	Pending
29	5	60	2024-05-29	621.18	Shipped	Pending
30	59	19	2024-12-25	922.9399999999999	Delivered	Paid
31	81	151	2024-05-15	1406.4	Delivered	Paid
32	15	140	2024-08-05	243.78	Placed	Paid
33	19	19	2024-05-16	696.46	Placed	Paid
34	77	192	2024-06-06	1050.24	Delivered	Pending
35	16	120	2024-08-12	138.0	Delivered	Pending
36	52	70	2024-12-24	649.02	Delivered	Pending
37	11	154	2024-08-12	68.62	Placed	Pending
38	78	65	2024-06-14	482.98	Placed	Paid
39	87	148	2024-04-27	1239.39	Delivered	Paid

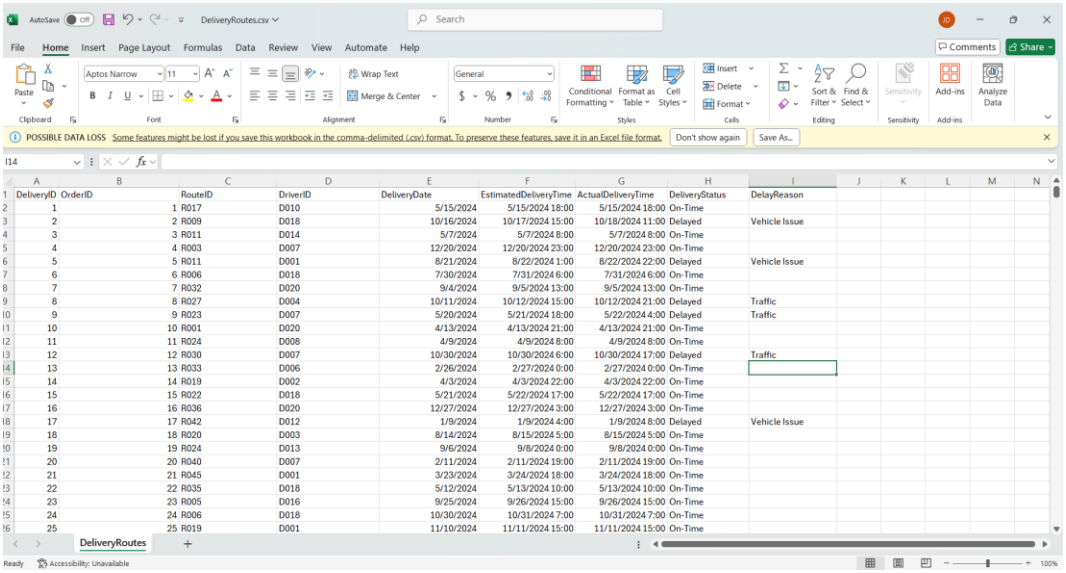
Figure 4: Orders.txt



OrderDetailID	OrderID	ProductID	Quantity	UnitPrice	LineTotal
1	1	20	37	8.44	312.28
2	1	14	15	2.86	42.9
3	1	137	73	5.2	379.6
4	2	147	55	6.32	347.6
5	2	37	8	6.91	55.28
6	2	16	7	4.51	31.57
7	2	24	42	4.1	172.2
8	3	59	55	3.01	165.55
9	3	18	98	4.71	461.58
10	3	131	32	8.16	261.12
11	3	100	28	1.6	44.8
12	3	57	20	9.35	187.0
13	4	85	97	8.87	860.39
14	4	29	42	2.57	107.94
15	4	153	34	7.78	264.52
16	4	21	30	2.08	62.4
17	5	35	20	1.26	25.2
18	5	81	82	8.9	729.8
19	5	38	38	4.34	164.92
20	5	11	21	6.9	144.9
21	6	95	7	4.29	30.03
22	7	61	55	3.09	169.95
23	7	108	66	5.44	359.04
24	8	64	4	5.3	21.2
25	8	95	32	4.29	137.28
26	8	153	17	7.78	132.26
27	8	131	82	8.16	669.12
28	8	83	16	9.29	148.64
29	9	90	32	4.13	132.16
30	10	59	81	3.01	243.81
31	10	151	97	4.78	463.66
32	11	97	54	6.59	355.86
33	12	27	85	6.79	577.15
34	12	87	91	7.93	721.63
35	12	150	71	7.56	536.76
36	12	62	15	2.5	37.5
37	13	17	59	4.89	288.51
38	13	69	97	6.55	635.35
39	13	69	47	6.55	307.85

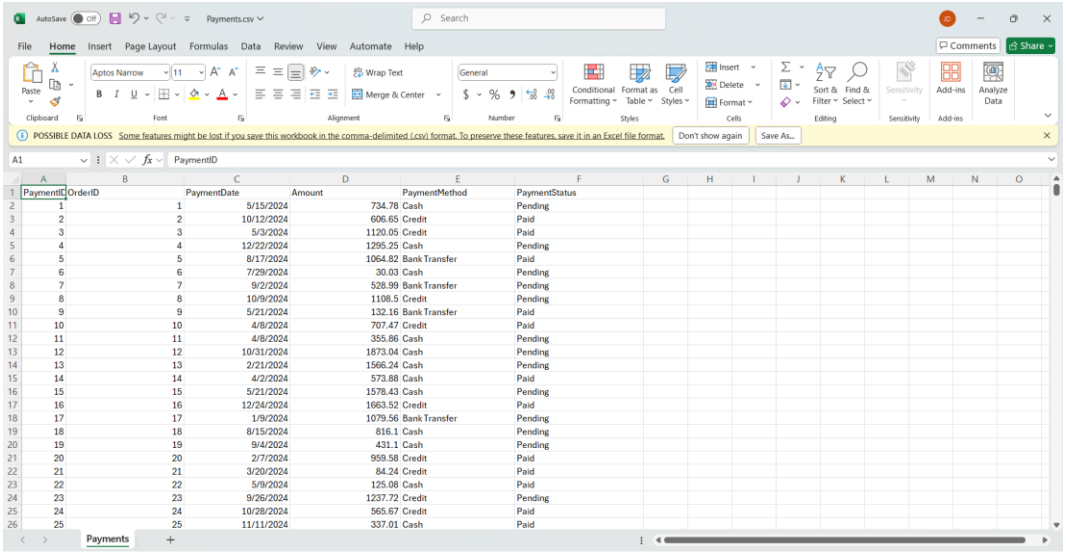
Ln 1, Col 1 4,435,309 characters 100% Windows (CRLF) UTF-8

Figure 5: OrderDetails.txt



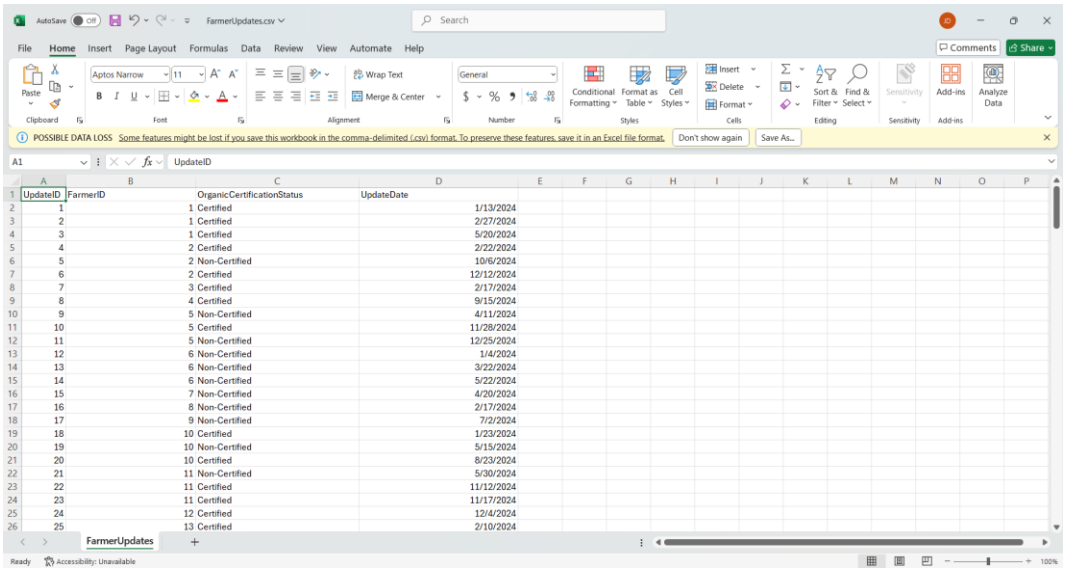
DeliveryID	OrderID	RouteID	DriverID	DeliveryDate	EstimatedDeliveryTime	ActualDeliveryTime	DeliveryStatus	DelayReason
1	1	R017	D010	5/15/2024	5/15/2024 18:00	5/15/2024 18:00	On-Time	
2	2	R009	D018	10/16/2024	10/17/2024 15:00	10/18/2024 11:00	Delayed	Vehicle Issue
3	3	R011	D014	5/7/2024	5/7/2024 8:00	5/7/2024 8:00	On-Time	
4	4	R003	D007	12/20/2024	12/20/2024 23:00	12/20/2024 23:00	On-Time	
5	5	R011	D001	8/21/2024	8/22/2024 1:00	8/22/2024 22:00	Delayed	Vehicle Issue
6	6	R006	D018	7/30/2024	7/31/2024 6:00	7/31/2024 6:00	On-Time	
7	7	R032	D020	9/4/2024	9/5/2024 13:00	9/5/2024 13:00	On-Time	
8	8	R027	D004	10/11/2024	10/12/2024 15:00	10/12/2024 21:00	Delayed	Traffic
9	9	R023	D007	5/20/2024	5/21/2024 18:00	5/22/2024 4:00	Delayed	Traffic
10	10	R001	D020	4/13/2024	4/13/2024 21:00	4/13/2024 21:00	On-Time	
11	11	R024	D008	4/9/2024	4/9/2024 8:00	4/9/2024 8:00	On-Time	
12	12	R030	D007	10/30/2024	10/30/2024 6:00	10/30/2024 17:00	Delayed	Traffic
13	13	R033	D006	2/26/2024	2/27/2024 0:00	2/27/2024 0:00	On-Time	
14	14	R019	D002	4/3/2024	4/3/2024 22:00	4/3/2024 22:00	On-Time	
15	15	R022	D018	5/21/2024	5/22/2024 17:00	5/22/2024 17:00	On-Time	
16	16	R036	D020	12/27/2024	12/27/2024 3:00	12/27/2024 3:00	On-Time	
17	17	R042	D012	1/9/2024	1/9/2024 4:00	1/9/2024 8:00	Delayed	Vehicle Issue
18	18	R020	D003	8/14/2024	8/15/2024 5:00	8/15/2024 5:00	On-Time	
19	19	R024	D013	9/6/2024	9/8/2024 0:00	9/8/2024 0:00	On-Time	
20	20	R040	D007	2/11/2024	2/11/2024 19:00	2/11/2024 19:00	On-Time	
21	21	R045	D001	3/23/2024	3/24/2024 18:00	3/24/2024 18:00	On-Time	
22	22	R035	D018	5/12/2024	5/13/2024 10:00	5/13/2024 10:00	On-Time	
23	23	R005	D016	9/25/2024	9/26/2024 15:00	9/26/2024 15:00	On-Time	
24	24	R006	D018	10/30/2024	10/31/2024 7:00	10/31/2024 7:00	On-Time	
25	25	R019	D001	11/10/2024	11/11/2024 15:00	11/11/2024 15:00	On-Time	

Figure 6: DeliveryRoutes.csv



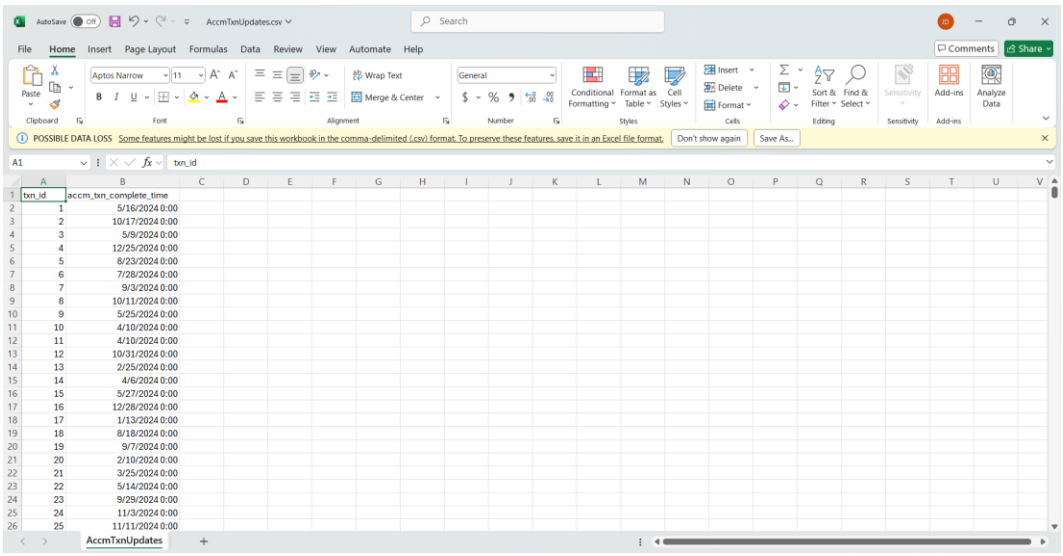
PaymentID	OrderID	PaymentDate	Amount	PaymentMethod	PaymentStatus
1	1	5/15/2024	734.78	Cash	Pending
2	2	10/12/2024	606.65	Credit	Paid
3	3	5/3/2024	1120.05	Credit	Paid
4	4	12/22/2024	1295.25	Cash	Pending
5	5	8/17/2024	1064.82	Bank Transfer	Paid
6	6	7/29/2024	30.03	Cash	Pending
7	7	9/2/2024	528.99	Bank Transfer	Pending
8	8	10/9/2024	1108.5	Credit	Pending
9	9	5/21/2024	132.16	Bank Transfer	Paid
10	10	4/8/2024	707.47	Credit	Paid
11	11	4/8/2024	355.86	Cash	Pending
12	12	10/31/2024	1873.04	Cash	Pending
13	13	2/21/2024	1566.24	Cash	Pending
14	14	4/2/2024	573.88	Cash	Paid
15	15	5/21/2024	1578.43	Cash	Pending
16	16	12/24/2024	1663.52	Credit	Paid
17	17	1/9/2024	1079.56	Bank Transfer	Pending
18	18	8/15/2024	816.1	Cash	Pending
19	19	9/4/2024	431.1	Cash	Pending
20	20	2/7/2024	959.58	Credit	Paid
21	21	3/20/2024	84.24	Credit	Paid
22	22	5/9/2024	125.08	Cash	Paid
23	23	9/26/2024	1237.72	Credit	Pending
24	24	10/28/2024	565.67	Credit	Paid
25	25	11/11/2024	337.01	Cash	Paid

Figure 7: Payments.csv



UpdateID	FarmerID	OrganicCertificationStatus	UpdateDate
1	1	Certified	1/13/2024
2	1	Certified	2/27/2024
3	1	Certified	5/20/2024
4	2	Certified	2/22/2024
5	2	Non-Certified	10/6/2024
6	2	Certified	12/12/2024
7	3	Certified	2/17/2024
8	4	Certified	9/15/2024
9	5	Non-Certified	4/11/2024
10	5	Certified	11/28/2024
11	5	Non-Certified	12/25/2024
12	6	Non-Certified	1/4/2024
13	6	Non-Certified	3/22/2024
14	6	Non-Certified	5/22/2024
15	7	Non-Certified	4/20/2024
16	8	Non-Certified	2/17/2024
17	9	Non-Certified	7/2/2024
18	10	Certified	1/23/2024
19	10	Non-Certified	5/15/2024
20	10	Certified	8/23/2024
21	11	Non-Certified	5/30/2024
22	11	Certified	11/12/2024
23	11	Certified	11/17/2024
24	12	Certified	12/4/2024
25	13	Certified	2/10/2024

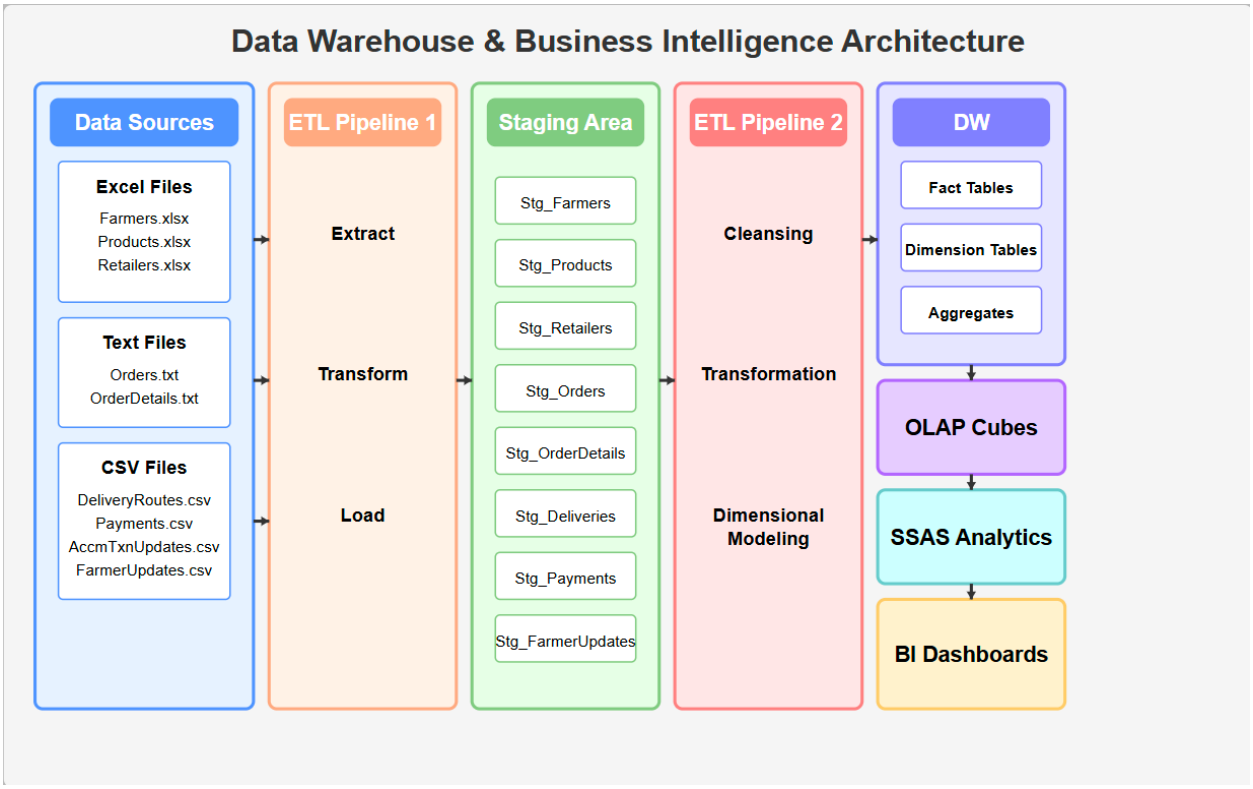
Figure 8: FarmerUpdates.csv



txn_id	accm_txn_complete_time
1	5/16/2024 0:00
2	10/17/2024 0:00
3	5/9/2024 0:00
4	12/25/2024 0:00
5	8/23/2024 0:00
6	7/28/2024 0:00
7	9/3/2024 0:00
8	10/11/2024 0:00
9	5/25/2024 0:00
10	4/10/2024 0:00
11	4/10/2024 0:00
12	10/31/2024 0:00
13	2/25/2024 0:00
14	4/6/2024 0:00
15	5/27/2024 0:00
16	12/28/2024 0:00
17	1/13/2024 0:00
18	8/18/2024 0:00
19	9/7/2024 0:00
20	2/10/2024 0:00
21	3/25/2024 0:00
22	5/14/2024 0:00
23	9/29/2024 0:00
24	11/3/2024 0:00
25	11/11/2024 0:00

Figure 9: AccmTxnUpdates.csv (Step 6)

Step 3: Solution architecture



1. **Data Sources:**
    - Summary: Heterogeneous files containing transactional and master data.
    - Details: Includes Farmers.xlsx (100 farmers), Products.xlsx (200 products), Retailers.xlsx (200 retailers), Orders.txt (50,000 orders), OrderDetails.txt (150,000 order details), DeliveryRoutes.csv (50,000 deliveries), Payments.csv (50,000 payments), AccmTxnUpdates.csv (50,000 updates, txn_id = OrderID), and synthetic **FarmerUpdates.csv** for OrganicCertificationStatus changes.
    - Role: Provides raw data for ETL.
2. **ETL Layer:**
    - Summary: Processes data extraction, transformation, and loading.
    - Details: Uses SSIS to extract from Excel, TXT, CSV, transform data (e.g., txn_process_time, ValidFrom for FarmerDim), and load into FarmToTableDW. Includes optional staging area.
    - Role: Integrates and cleanses data, populating fact and dimension tables.
3. **Data Warehouse Layer:**
    - Summary: SQL Server database storing data in a star schema.
    - Details: FarmToTableDW includes the central fact table OrderSalesFact, which holds key measures such as TotalAmount, Quantity, and order-level details including Status, PaymentStatus, and PaymentMethod. Payment-related attributes have been consolidated into this fact table instead of a separate PaymentFact. The warehouse is supported by five dimension tables: FarmerDim (implemented with Type 1 Slowly Changing Dimensions using ETL-generated), ProductDim, RetailerDim, DateDim, and DeliveryDim.
    - Role: Stores data for analytical processing.
4. **Analytical Layer:**
    - Summary: Processes multidimensional queries via SSAS cubes.
    - Details: Aggregates measures (TotalAmount) across dimensions (e.g., Category, OrganicCertificationStatus) for fast queries.
    - Role: Enables efficient analysis for reports.

5. **Presentation Layer:**
    - Summary: Delivers visualizations and reports.
    - Details: Uses Power BI/SSRS to create dashboards for eight reports, including charts and tables.
    - Role: Provides actionable insights to stakeholders

Step 4: Data warehouse design & development

Dimensional Model Description

The data warehouse for the Farm-to-Table Produce Delivery Tracker dataset is designed as a star schema, comprising two fact tables (OrderSalesFact, PaymentFact) and five-dimension tables (FarmerDim, ProductDim, RetailerDim, DateDim, DeliveryDim). The scheme supports eight analytical reports: Sales by Product Category, Delivery Efficiency, Farmer Performance, Retailer Sales by Type, Top Products by Revenue, Regional Sales Distribution, Organic Certification Impact, and Payment Method Trends.

- **Fact Table: OrderSalesFact**
  The OrderSalesFact table captures comprehensive order-level and payment-level transactional data to support sales, performance, and payment trend analysis. It includes key measures such as TotalAmount (revenue) and Quantity (items ordered), along with transactional attributes like Status, PaymentStatus, and PaymentMethod (a degenerate dimension). The txn_process_time is assumed to be derived or can be inferred based on order processing logic.

  Foreign keys including FarmerSK, ProductSK, RetailerSK, DateKey, and DeliverySK connect to the respective dimension tables (FarmerDim, ProductDim, RetailerDim, DateDim, DeliveryDim).
  The table is populated from multiple sources:

  Orders.txt and **OrderDetails.txt** (for order data),

  AccmTxnUpdates.csv (mapped via txn_id = OrderID for transaction status and updates),

  DeliveryRoutes.csv (for delivery-related information).

  Payment-related information (e.g., PaymentMethod, PaymentStatus, PaymentID) is embedded directly in the OrderSalesFact to simplify design, since PaymentMethod has low cardinality and does not require a separate dimension table.

- **Dimension Tables:**
  - **FarmerDim:** Includes FarmerSK (surrogate key), SourceFarmerID (business key), Name, FarmName, City, State, Zip, OrganicCertificationStatus,

FarmSize, and Type 1 SCD attributes (). Since Farmers.xlsx provides a static snapshot, ValidFrom, ValidTo, and IsCurrent are generated during ETL (Step 5) using a synthetic FarmerUpdates.csv or assumed static with scalable SCD logic. Supports Organic Certification Impact and Farmer Performance.

  - **ProductDim:** Includes ProductID, ProductName, Category, SubCategory, UnitPrice, SeasonalAvailability. Supports Sales by Product Category and Top Products.
  - **RetailerDim**: Includes RetailerID, RetailerName, Type, City, State, Zip. Supports Retailer Sales and Regional Sales.
  - **DateDim:** Includes DateID, Date, Year, Quarter, Month, Day. Supports time-based analysis.
  - **DeliveryDim:** Includes DeliveryID, OrderID, RouteID, DriverID, DeliveryDate, DeliveryStatus, DelayReason. Supports Delivery Efficiency.

- **Slowly Changing Dimension:** FarmerDim implements a **Type 1 Slowly Changing Dimension (SCD)** for attributes such as OrganicCertificationStatus, City, FarmName, FarmSize, and Name. Changes to these fields overwrite previous values without retaining history, ensuring current-state reporting.

  ProductDim uses **Type 1 SCD** logic for attributes like SeasonalAvailability and UnitPrice, supporting up-to-date reporting for product availability and pricing without maintaining historical records.

  DeliveryDim also applies **Type 1 SCD** to track updates in delivery-related attributes such as ActualDeliveryTime, DeliveryDate, and EstimatedDelivery, ensuring operational reporting reflects the most recent delivery details.

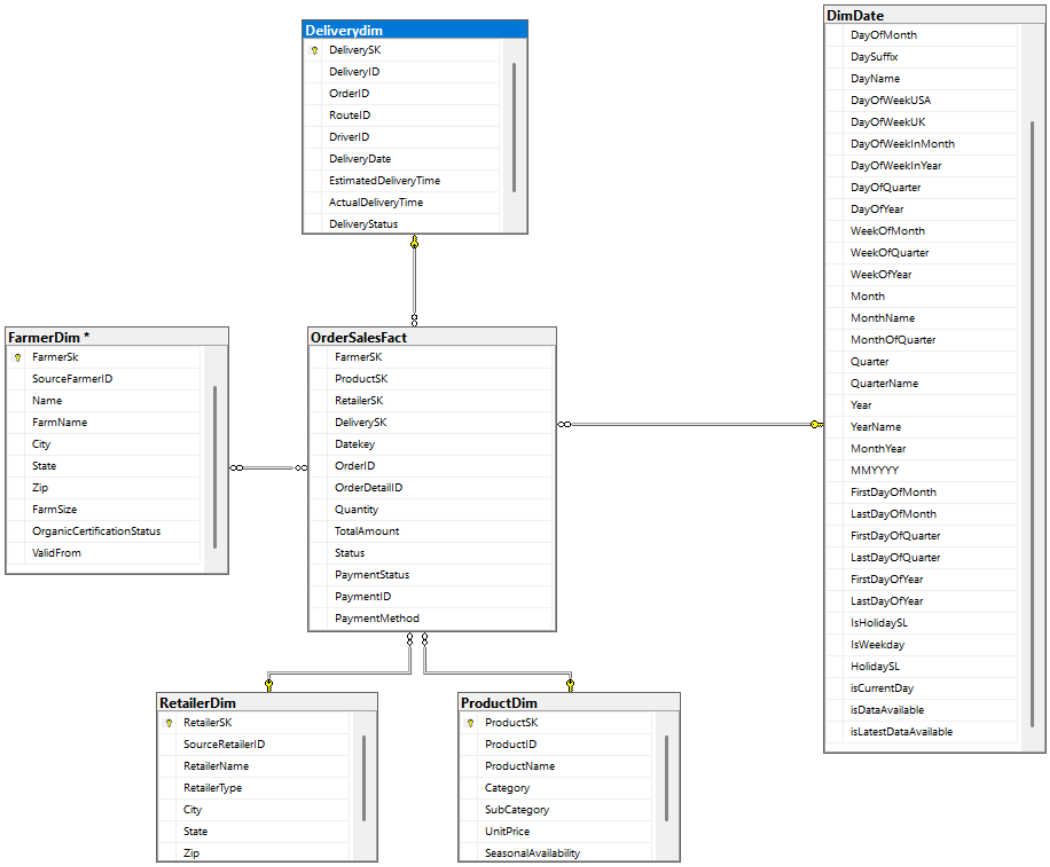
  This simplified Type 1 approach ensures easier ETL management and performance, suitable when historical tracking is not a business requirement.

- **PaymentMethod Emphasis:** PaymentMethod is a **degenerate dimension** included directly in the OrderSalesFact table to support payment method trend analysis (e.g., pie charts showing distribution of methods like Credit Card, Bank Transfer, or Cash). It is a non-numeric VARCHAR field sourced from Payments.csv and retained within the fact table due to its **low cardinality** and lack of additional descriptive attributes. This design choice avoids the need for a separate PaymentMethodDim table, simplifying the schema while fully supporting reporting requirements.

Assumptions

1. OrderSalesFact is the single central fact table capturing both **order-level and payment-level** transactional data at the order granularity.
2. PaymentMethod is a **degenerate dimension** stored directly in OrderSalesFact, sourced from Payments.csv, with no separate PaymentMethodDim table due to low cardinality and no additional attributes.
3. **Type 1 Slowly Changing Dimension (SCD)** logic is applied to:
    - FarmerDim for attributes: OrganicCertificationStatus, City, FarmName, FarmSize, and Name
    - ProductDim for: SeasonalAvailability, UnitPrice
    - DeliveryDim for: ActualDeliveryTime, DeliveryDate, and EstimatedDelivery
      These attributes are overwritten on change, with no historical tracking retained.
4. DateDim covers the year 2024, based on OrderDate and PaymentDate.
5. Dataset maintains referential integrity, with txn_id matching OrderID.

Relational Diagram



Step 5: ETL development

The ETL (Extract, Transform, Load) process is designed in two pipelines, moving data from raw source files (Excel, Text, CSV) through a staging area and into the final dimensional model of the data warehouse (DW), which supports OLAP cubes, SSAS analytics, and BI dashboards.

ETL Pipeline 1: Extract → Transform → Load to Staging Area

Step 1: Data Extraction
Source files are extracted from multiple formats:

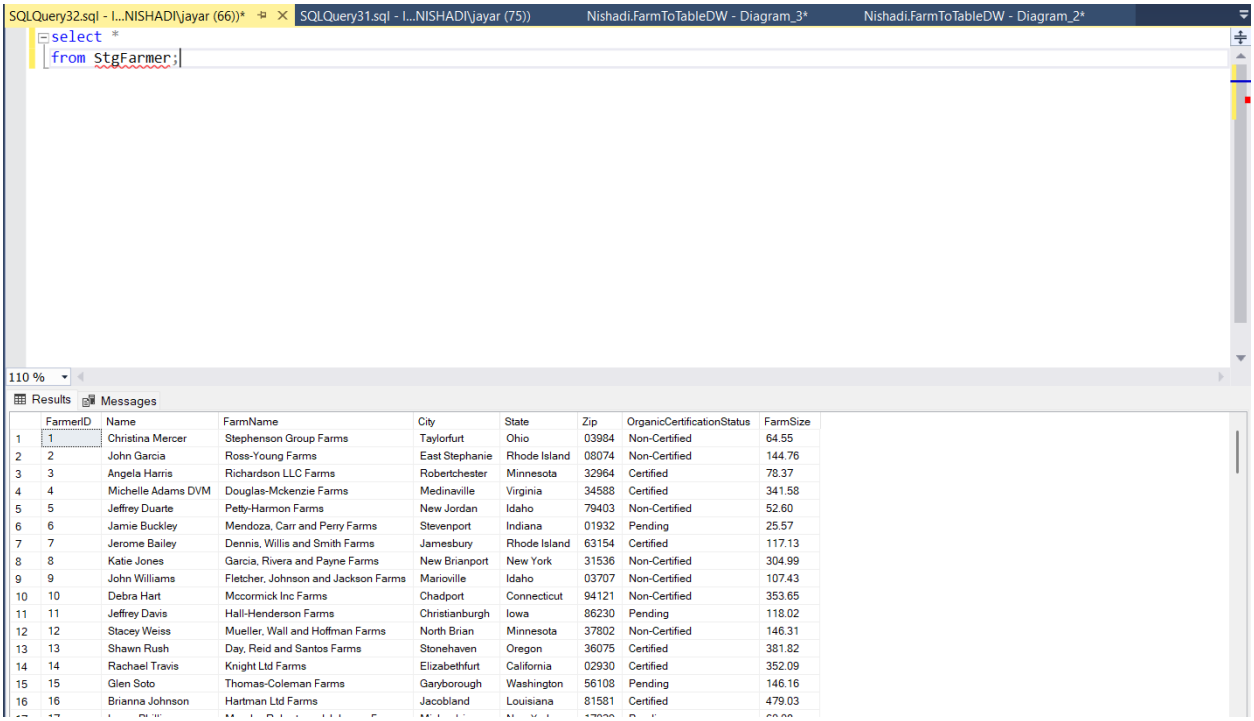
- **Excel Files**: Farmers.xlsx, Products.xlsx, Retailers.xlsx
- **Text Files**: Orders.txt, OrderDetails.txt
- **CSV Files**: DeliveryRoutes.csv, Payments.csv, AccmTxnUpdates.csv, FarmerUpdates.csv

Step 2: Transformation (Initial Parsing and Standardization)

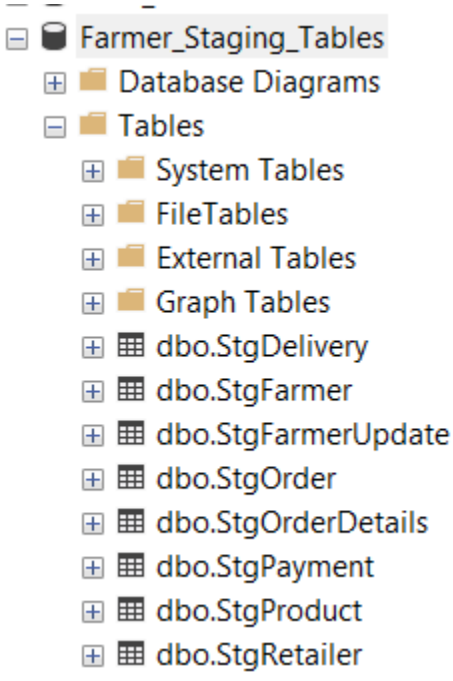
- Basic parsing, type conversions, column renaming, and date standardization.
- Ensure keys like FarmerID, RetailerID, OrderID, txn_id, and DeliveryID are aligned and validated.

Step 3: Load to Staging Area
Transformed data is loaded into the staging tables:

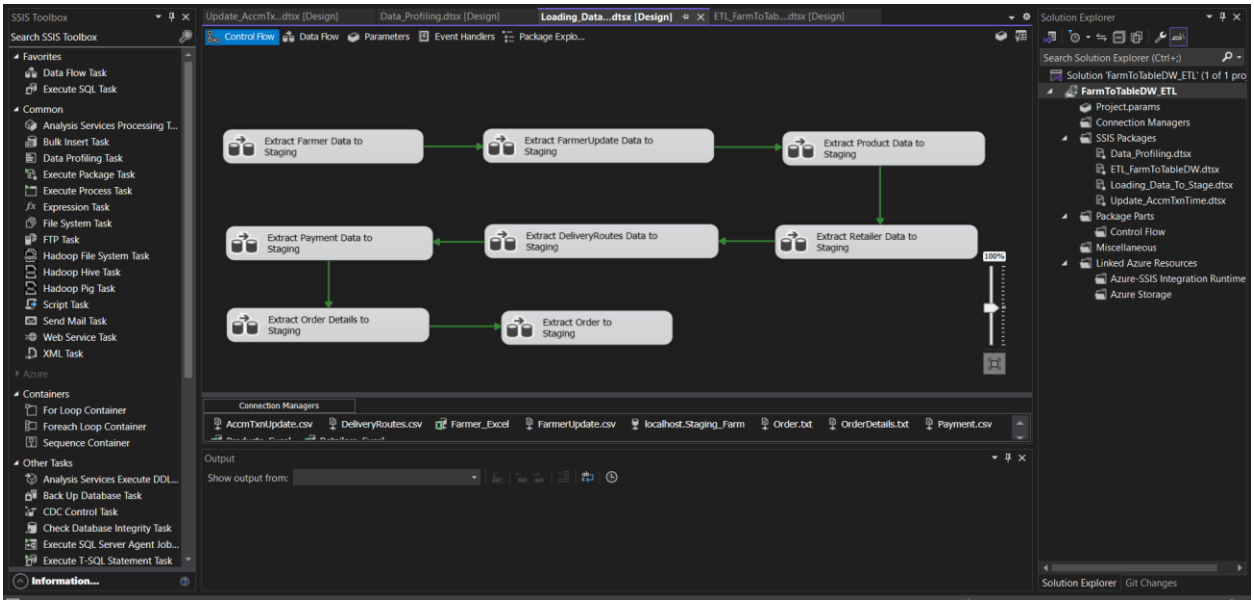
- Stg_Farmers
- Stg_Products
- Stg_Retailers
- Stg_Orders
- Stg_OrderDetails
- Stg_Deliveries
- Stg_Payments
- Stg_FarmerUpdates



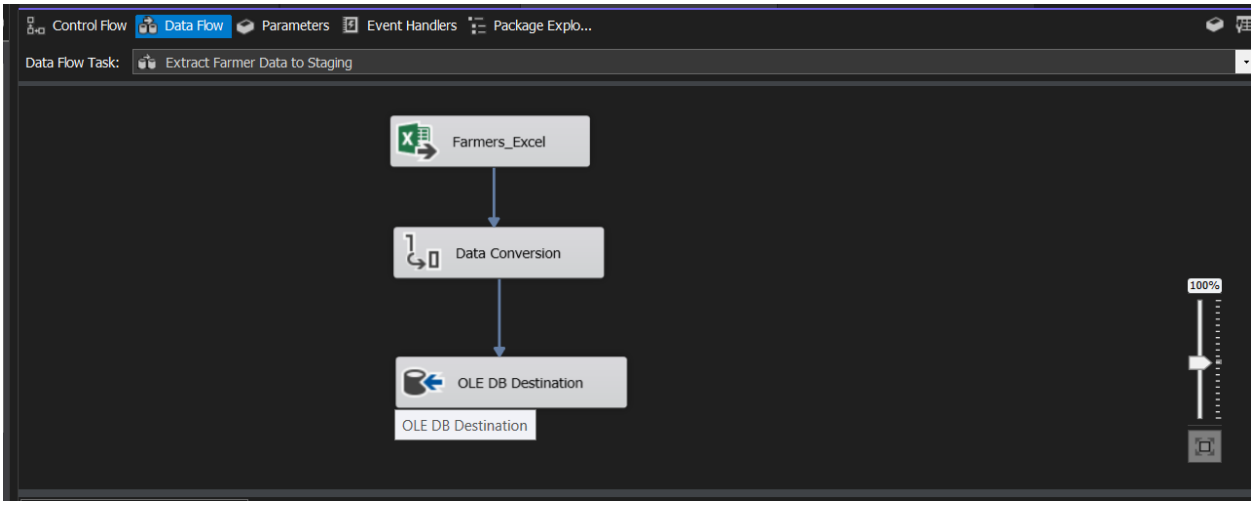
	FarmerID	Name	FarmName	City	State	Zip	OrganicCertificationStatus	FarmSize
1	1	Christina Mercer	Stephenson Group Farms	Taylorfurt	Ohio	03984	Non-Certified	64.55
2	2	John Garcia	Ross-Young Farms	East Stephanie	Rhode Island	08074	Non-Certified	144.76
3	3	Angela Harris	Richardson LLC Farms	Robertchester	Minnesota	32964	Certified	78.37
4	4	Michelle Adams DVM	Douglas-Mckenzie Farms	Medinaville	Virginia	34588	Certified	341.58
5	5	Jeffrey Duarte	Petty-Harmon Farms	New Jordan	Idaho	79403	Non-Certified	52.60
6	6	Jamie Buckley	Mendoza, Carr and Perry Farms	Stevenport	Indiana	01932	Pending	25.57
7	7	Jerome Bailey	Dennis, Willis and Smith Farms	Jamesbury	Rhode Island	63154	Certified	117.13
8	8	Katie Jones	Garcia, Rivera and Payne Farms	New Brianport	New York	31536	Non-Certified	304.99
9	9	John Williams	Fletcher, Johnson and Jackson Farms	Marioville	Idaho	03707	Non-Certified	107.43
10	10	Debra Hart	Mccormick Inc Farms	Chadport	Connecticut	94121	Non-Certified	353.65
11	11	Jeffrey Davis	Hall-Henderson Farms	Christianburgh	Iowa	86230	Pending	118.02
12	12	Stacey Weiss	Mueller, Wall and Hoffman Farms	North Brian	Minnesota	37802	Non-Certified	146.31
13	13	Shawn Rush	Day, Reid and Santos Farms	Stonehaven	Oregon	36075	Certified	381.82
14	14	Rachael Travis	Knight Ltd Farms	Elizabethfurt	California	02930	Certified	352.09
15	15	Glen Soto	Thomas-Coleman Farms	Garyborough	Washington	56108	Pending	146.16
16	16	Brianna Johnson	Hartman Ltd Farms	Jacobland	Louisiana	81581	Certified	479.03



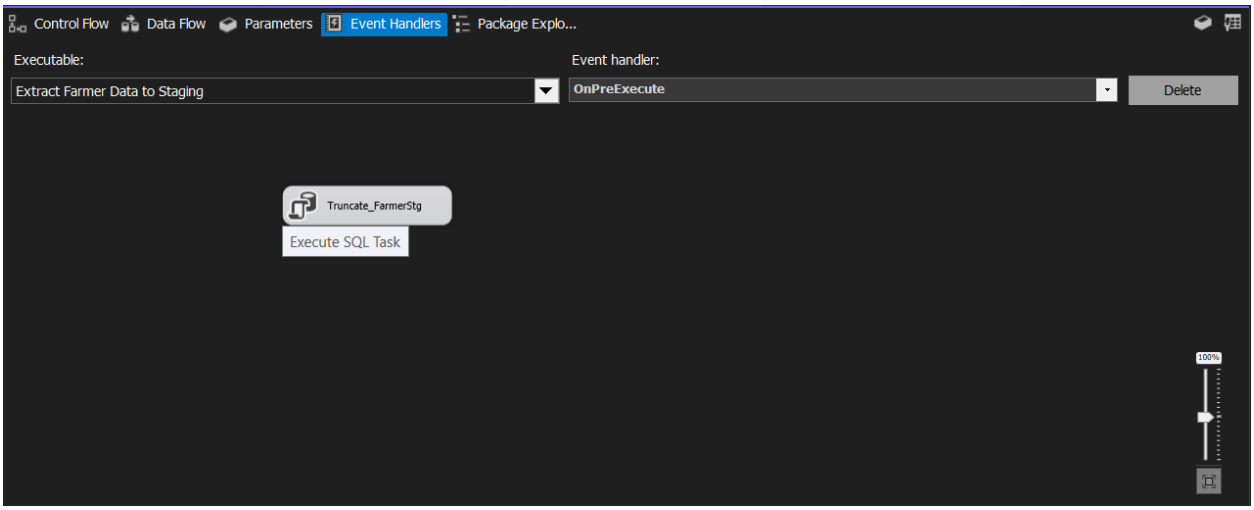
SSIS Control Flow (Staging Table)



Data Flow: Flat file



Event Handler to Stop duplicate data.



ETL Pipeline 2: Cleansing → Transformation → Dimensional Modeling → Load to DW

Step 4: Data Cleansing

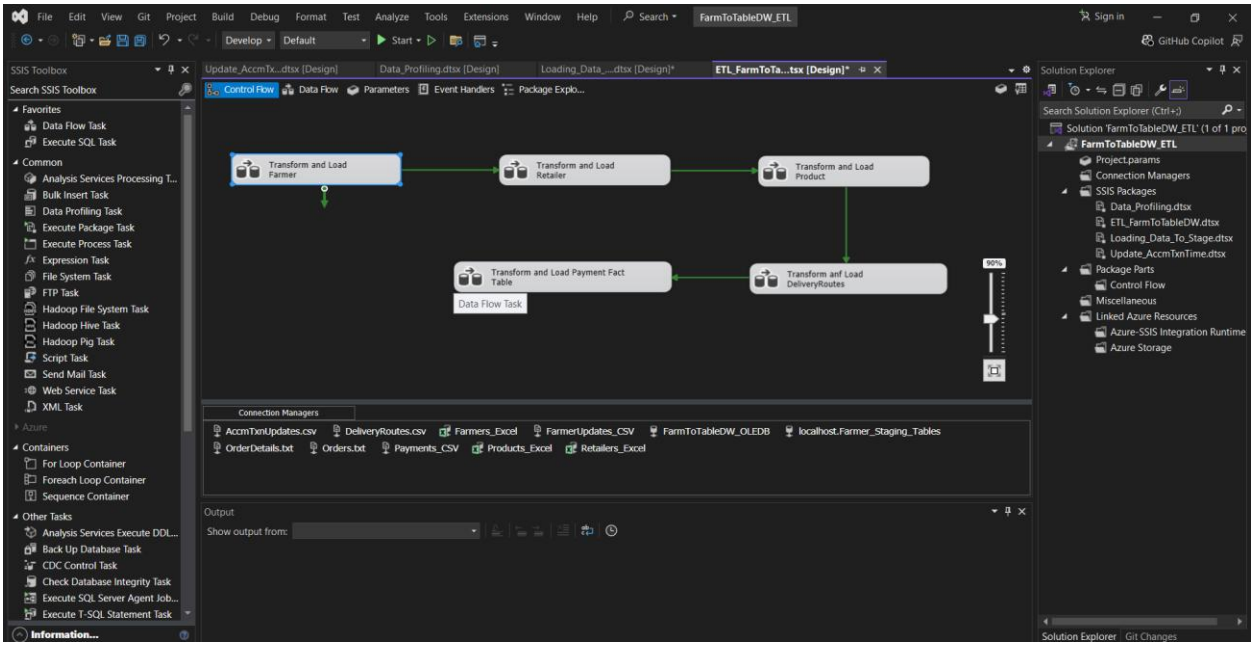
- Deduplication, null handling, and referential integrity checks.
- Ensure joins between staging tables (e.g., Stg_OrderDetails with Stg_Orders, Stg_Payments with Stg_Orders using OrderID).

Step 5: Transformation (Business Logic Application)

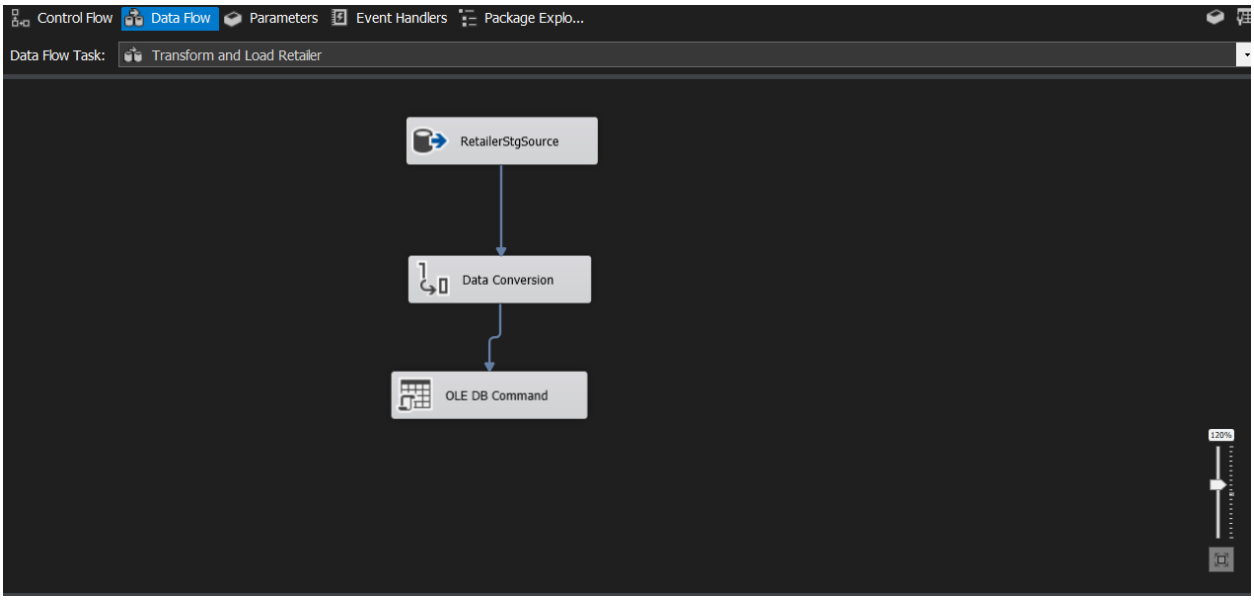
- **Dimensional Attributes Update:**
    - Apply **Type 1 SCD** logic to dimension attributes:
        - FarmerDim: OrganicCertificationStatus, City, FarmName, FarmSize, Name
        - ProductDim: SeasonalAvailability, UnitPrice
        - DeliveryDim: ActualDeliveryTime, DeliveryDate, EstimatedDelivery
- Combine Orders, OrderDetails and Payments to form OrderSalesFact.
- Extract PaymentMethod from Payments.csv as a **degenerate dimension** directly in OrderSalesFact.

Step 6: Dimensional Modeling and Load to DW

- Populate **Dimension Tables**:
    - FarmerDim, ProductDim, RetailerDim, DateDim, DeliveryDim
- Populated **Fact Table**:
    - OrderSalesFact, including both order and payment measures and PaymentMethod
- Ensure DateDim is generated to cover the full year of 2024.



Load Retailer Dimension using procedure.

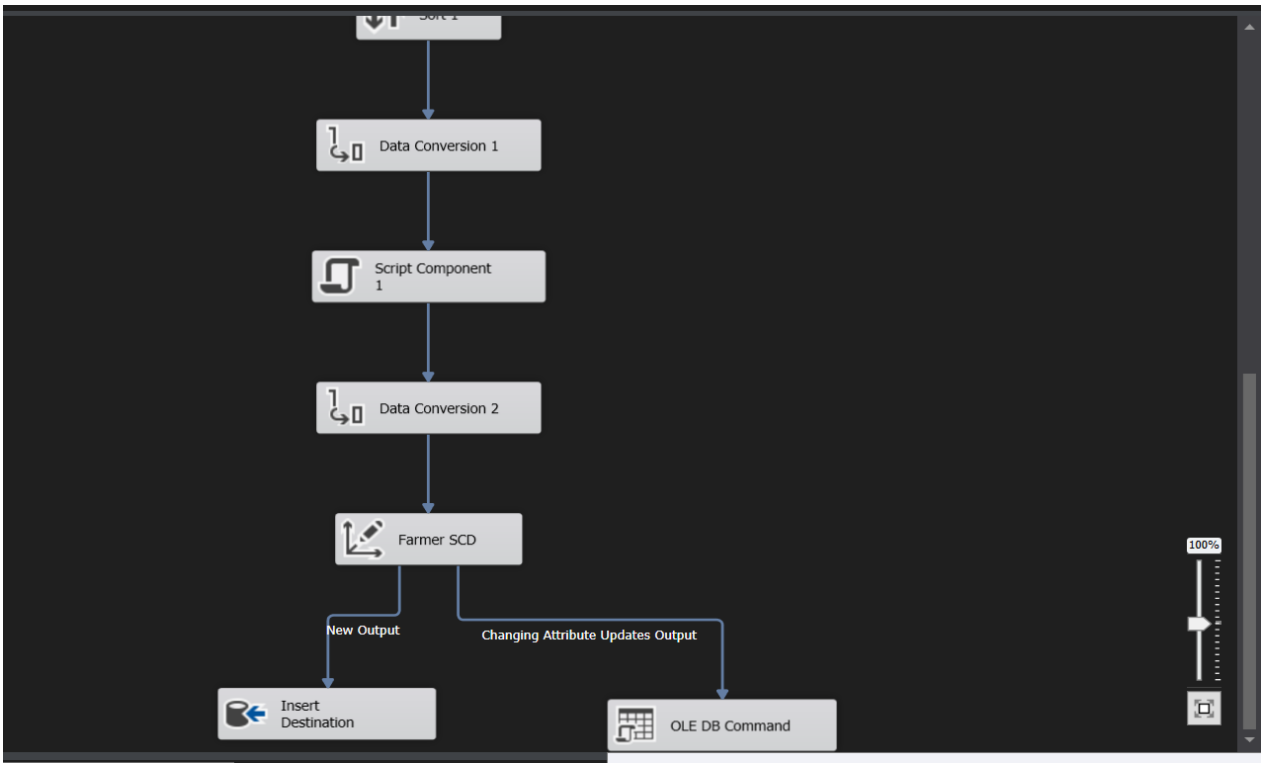


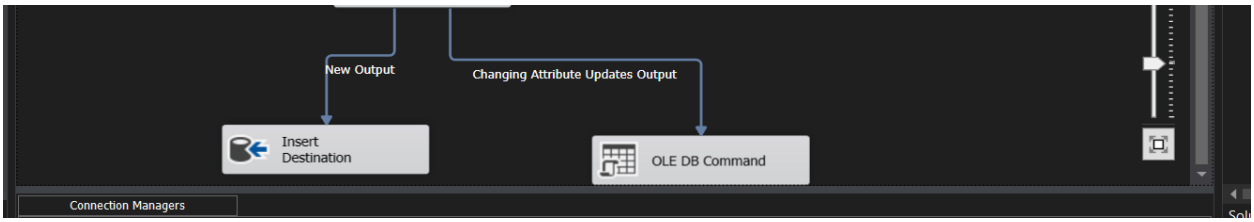

```
-- Drop existing stored procedure (if it exists)
IF OBJECT_ID('dbo.UpdateRetailerDim', 'P') IS NOT NULL
   DROP PROCEDURE dbo.UpdateRetailerDim;
GO

-- Create stored procedure
CREATE PROCEDURE dbo.UpdateRetailerDim
   @SourceRetailerID VARCHAR(10),
   @RetailerName VARCHAR(100),
   @RetailerType VARCHAR(50),
   @City VARCHAR(50),
   @State VARCHAR(50),
   @Zip VARCHAR(10)
AS
BEGIN
```

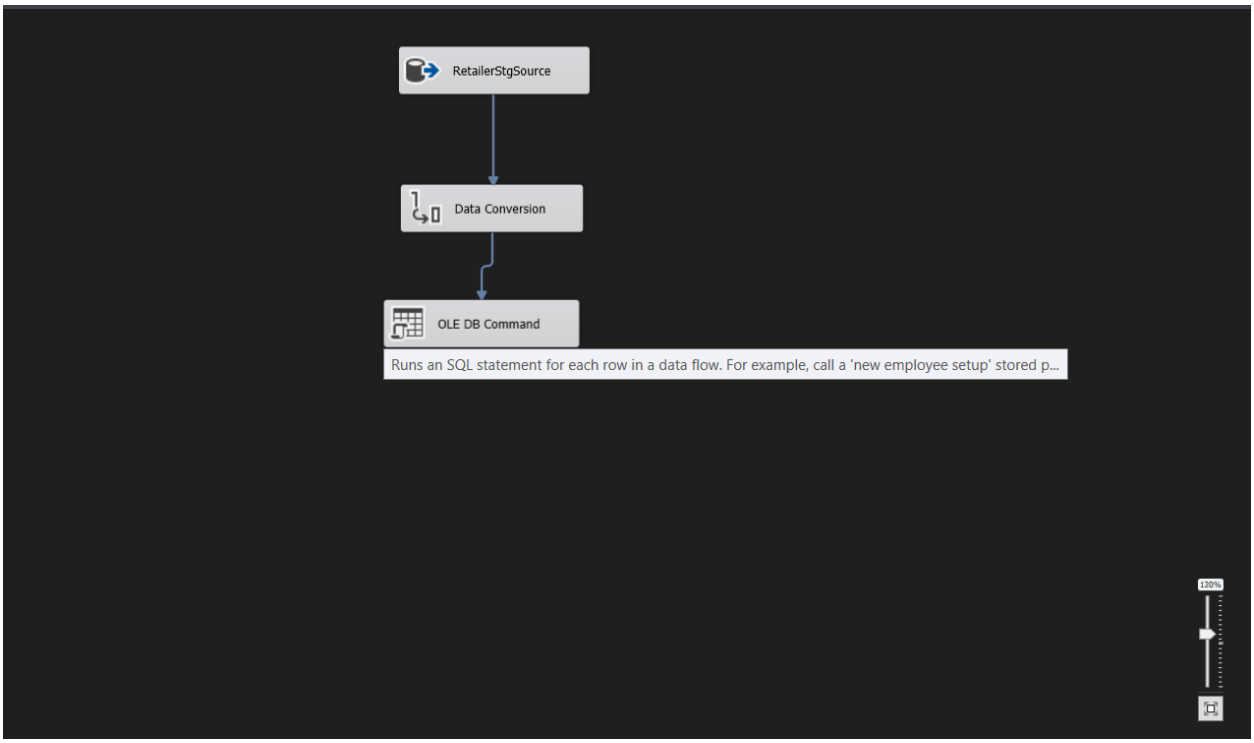
```
    -- Insert new record if SourceRetailerID doesn't exist
    IF NOT EXISTS (
        SELECT RetailerSK
        FROM dbo.RetailerDim
        WHERE SourceRetailerID = @SourceRetailerID
    )
    BEGIN
        INSERT INTO dbo.RetailerDim
            (SourceRetailerID, RetailerName, RetailerType, City, State, Zip)
        VALUES
            (@SourceRetailerID, @RetailerName, @RetailerType, @City, @State, @Zip);
    END
    ELSE
    BEGIN
        -- Update existing record
        UPDATE dbo.RetailerDim
        SET RetailerName = @RetailerName,
            RetailerType = @RetailerType,
            City = @City,
            State = @State,
            Zip = @Zip
        WHERE SourceRetailerID = @SourceRetailerID;
    END
END;
GO
```

1. Farmer Dimension:

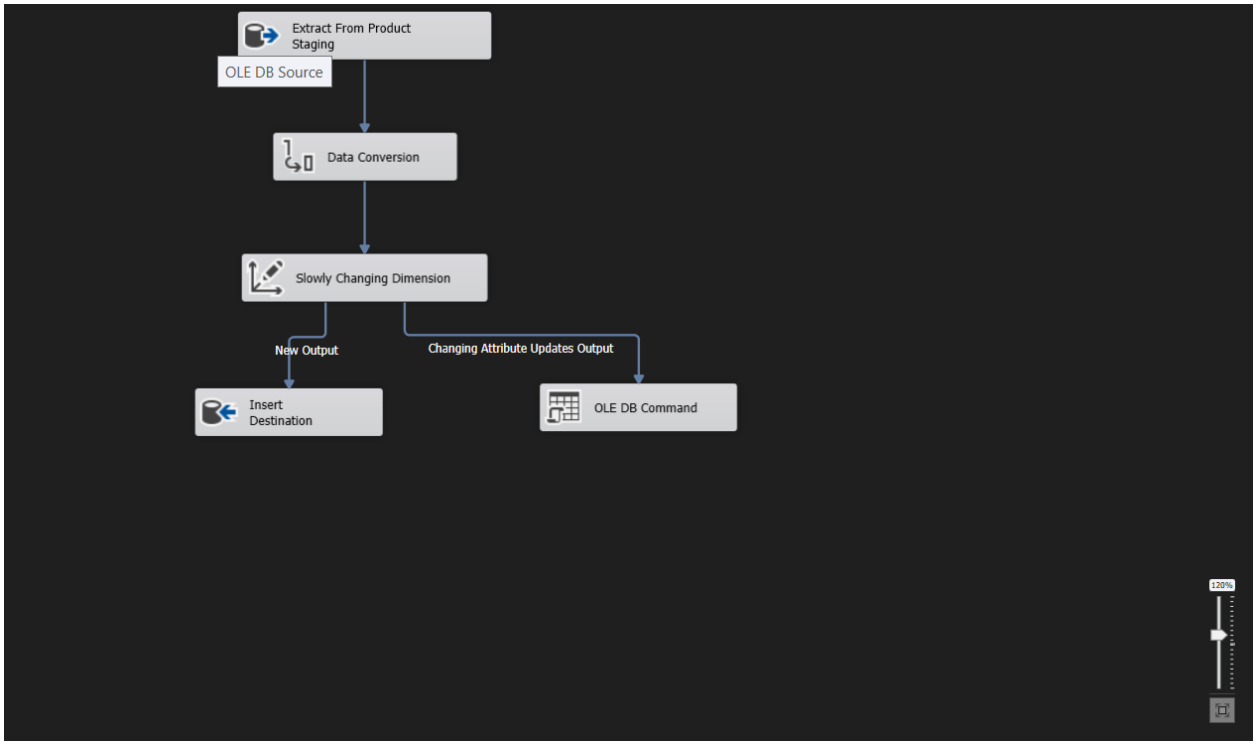




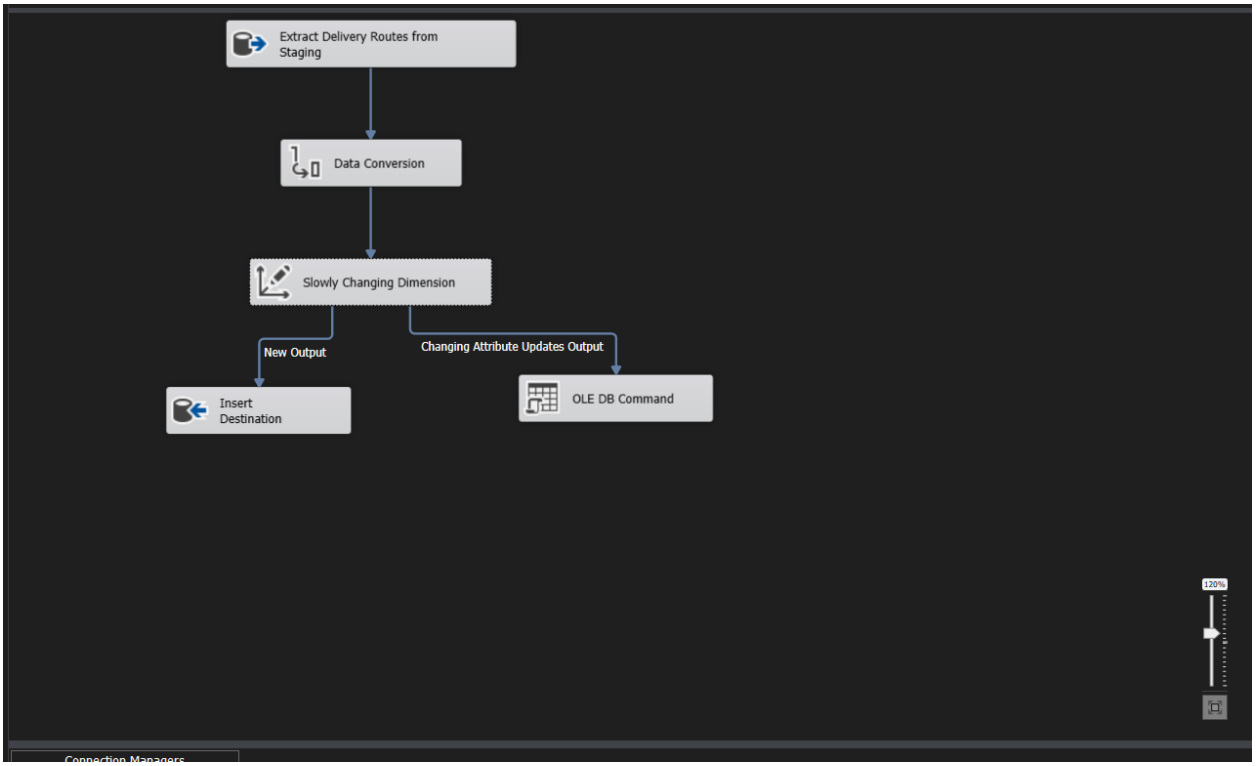
2. Retailer Dimension:



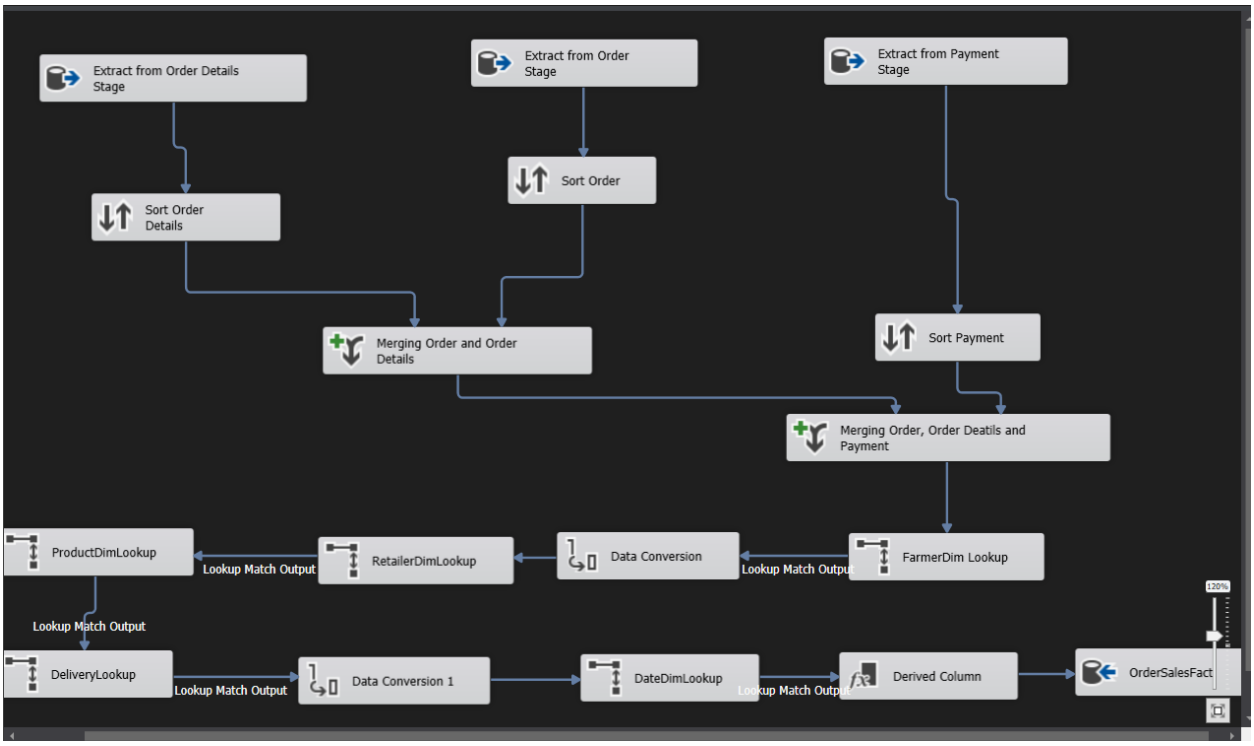
3. Product Dimension:



4. DeliveryRotes Dimension



5. Load Data to OrderSales Fact Table:



Step 6: ETL development – Accumulating fact tables