Crop Production

Report

Subject

ETL & Data Warehousing

Data ScienceTech Institute

Instructor

Nelson Lopez

Submitted by

Sofiane Chaoui , Shikhar Saini, Harsha K. S. Mudiyanselage

DSTI School of Engineering
20th april 2025

Table of contents

Abstract

This project implements an ETL and data warehousing solution using Microsoft Visual Studio and SSIS, based on the Crop Production in East Asia dataset from Kaggle. The goal is to transform raw agricultural data into a structured format for analysis. A star schema was designed with one fact table and four dimension tables—crop, country, year, and unit—to support efficient querying. The ETL process includes extracting data from CSV files, applying transformations, and loading it into a SQL Server data warehouse. The resulting system enables insights into crop production trends across East Asian countries over time.

1. Dataset

This dataset, published by the Ministry of Agriculture in India and compiled by the IMT Kaggle team, focuses on crop production data in East Asia over multiple years. It provides historical insights into how different crops performed across countries, which can be useful for trend analysis, policymaking, and agricultural planning.

What dataset includes:

- Crop types
- Countries
- Production volumes
- Units of measurement
- Years

The dataset is valuable for studying agricultural trends, forecasting production, and understanding how different regions focus on specific crops.

The main file relevant to your project are:

- Production_crop_E_Africa
- Production_crop_E_Asia

- Production_crop_E_America
- Production_crop_E_Europe
- Production_crop_E_Oceania

Key Columns in the dataset :

Column	Description
`Area`	Name of the country (e.g., China, Japan, Korea, etc.)
`Item`	Type of crop (e.g., Rice, Wheat, Maize)
`Element`	Type of data being measured (most commonly "Production")
`Unit`	Unit of measurement (e.g., tonnes)
`Year`	Year of record
`Value`	Amount of crop produced (numeric)
`Flag`	Optional metadata (could include info on estimation or data accuracy)

2. Star Schema Design

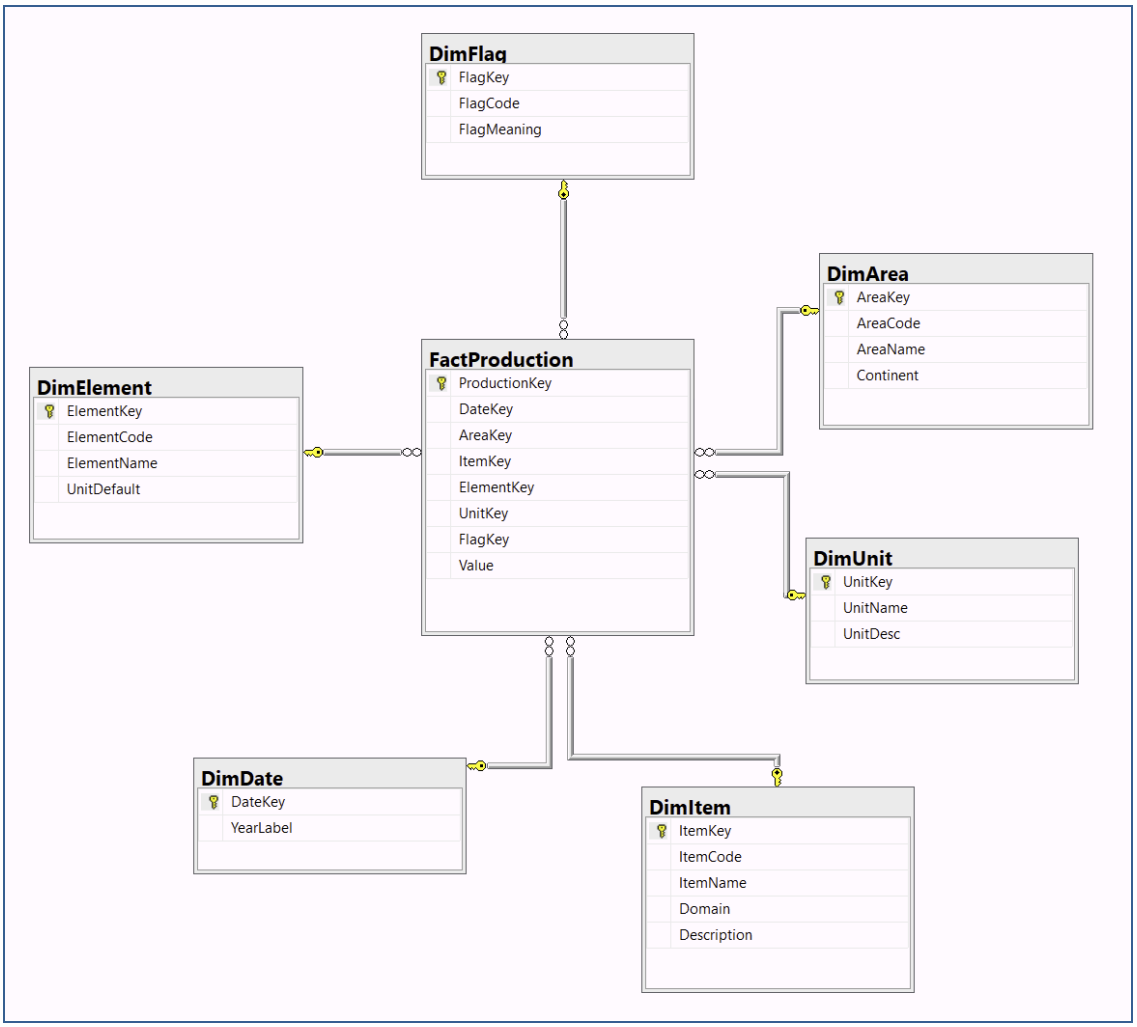
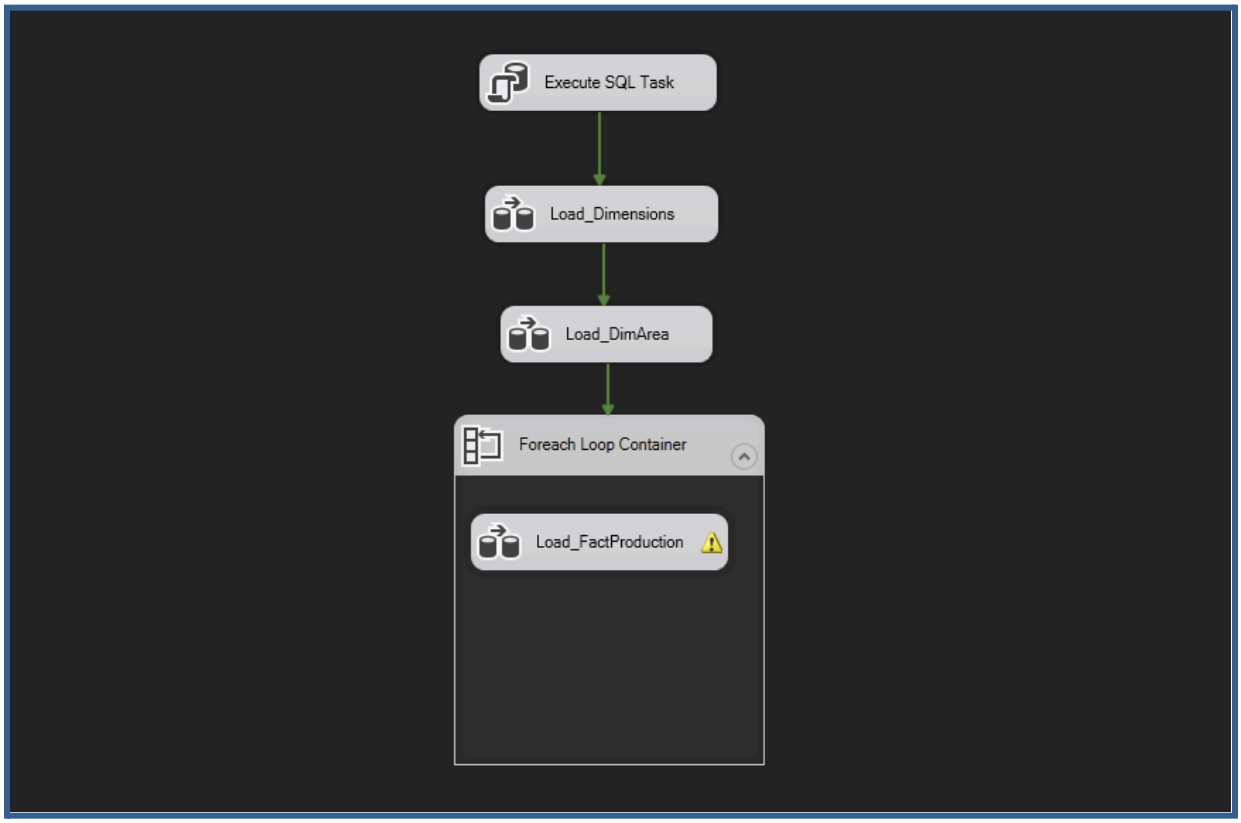


Table	Comment
FactProduction	1392960 rows after full load
DimDate	generated once by script
DimArea	deduped from production files
DimItem	from items.csv
DimElement	from elements.csv
DimUnit	from units.csv
DimFlag	from flags.csv

3. ETL Process and key transformations

a. ETL control flow architecture



Order	Task	What it does
1	Execute SQL	Generates 59 rows in DimDate with a simple loop
2	Load_Dimensions	Four parallel pipelines load Item, Element, Unit, Flag
3	Load_DimArea	One pipeline per continent, aggregates distinct
4	Foreach Loop -> Load_FactProduction	Re-uses the same Data Flow for the five production files

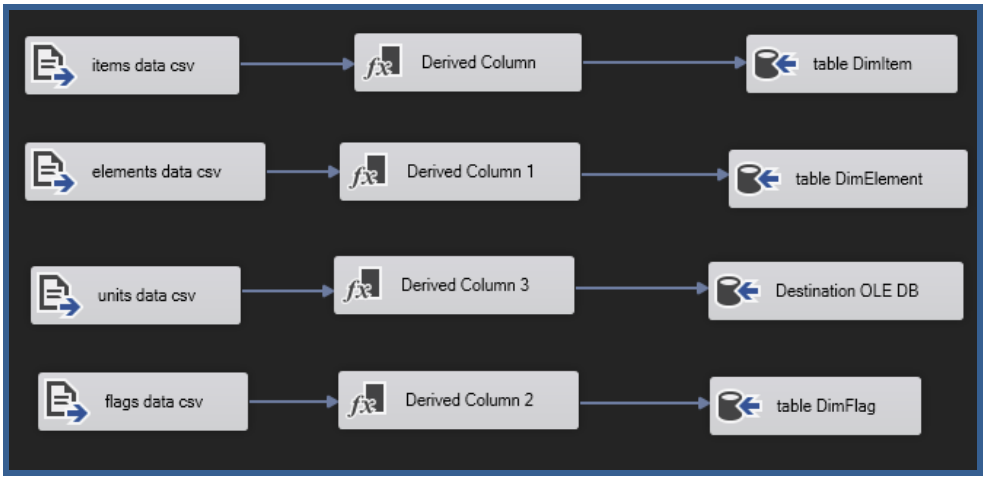
b. Data-cleaning techniques

Problem	Transformation & expression
Embedded quotes	REPLACE(MyColumn,"\"","") in **Derived Column**
Trailing blanks	TRIM(Column)
Mixed case units	UnitUp = UPPER(Unit)
Wide year columns	Unpivot ⟶ YearLabel + RawValue
Blank or zero values	Conditional Split (`ISNULL(RawValue)
Unmatched business keys	Lookups set to Redirect row; bad rows logged, package continues
Merge value & flag streams	Sort (AreaCode, ItemCode, ElementCode, DateKey) + Merge Join

**keeping transformations in SSIS avoids staging tables, reduces I/O and makes the package portable to any SQL instance*

c. Key Data-Flows

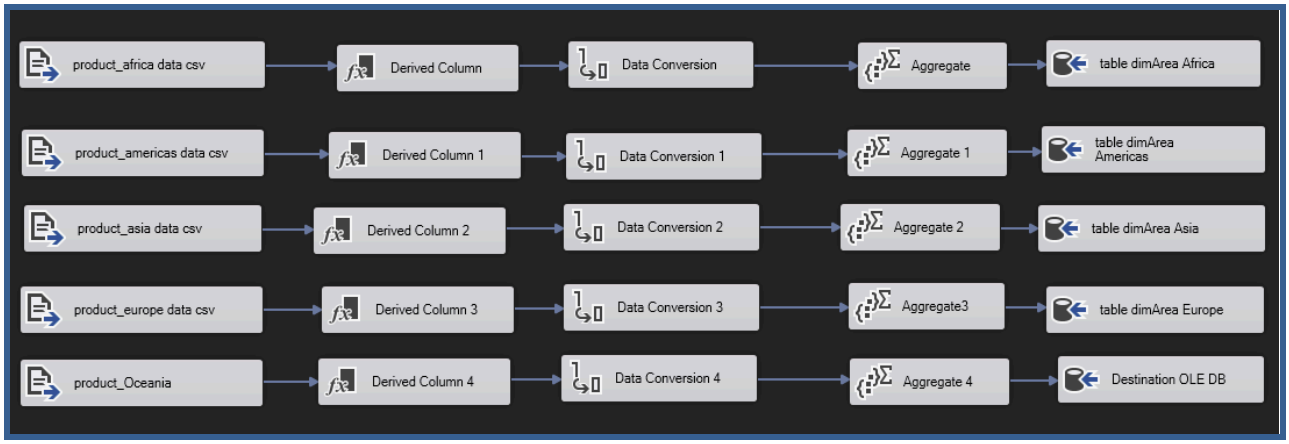
Load_Dimensions



** sources files : items.csv, elements.csv, units.csv, flags.csv*
** target table : DimItem, DimElement, DimUnit, DimFlag*

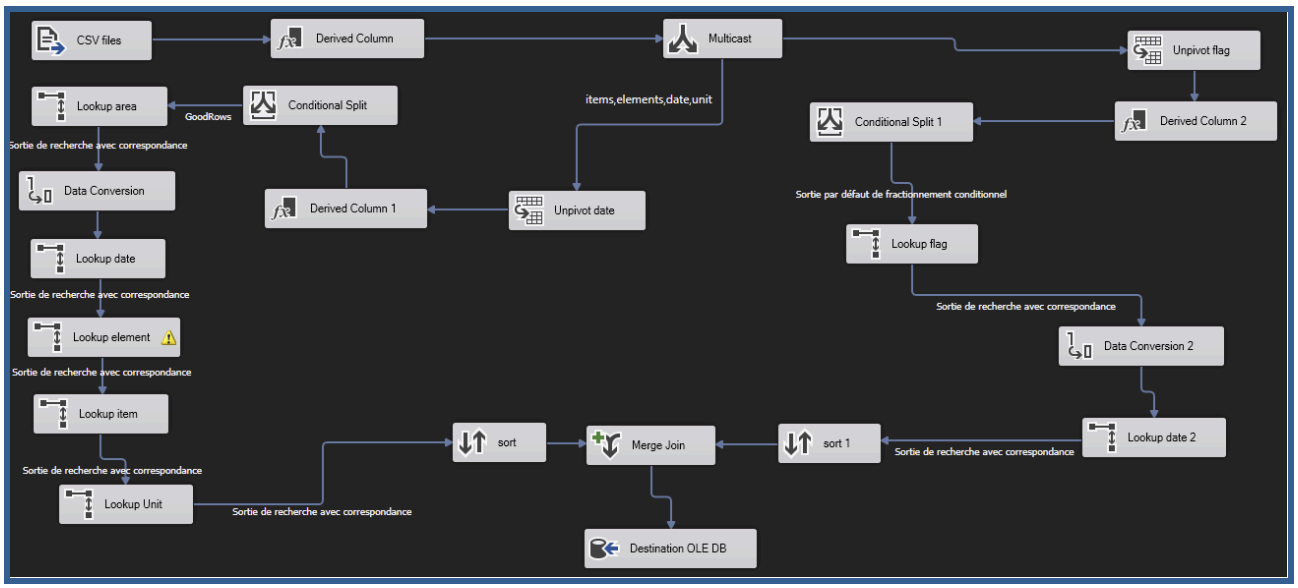
Each reference CSV is trimmed (REPLACE / TRIM) then fast-loaded as-is into its dimension; no business logic because these lists define the business keys used later, so we must preserve them

Load_DimArea



** sources files : Five production files Production_Crops_E_**
** target table : DimArea*
We read only AreaCode & AreaName, stamp the Continent literal that corresponds to the file, cast to Unicode, GROUP BY to drop duplicates, then insert; this guarantees one row per country and keeps continent in a single place for slicing queries.

Load_FactProduction



** sources files : same five Production_Crops_E_**
** target table : FactProduction*

- cleans quotes & upper-cases units,
- Unpivots 59 year columns to atomic rows,
- filters NULL/0 to shrink fact size,
- looks up every surrogate key (Date, Area, Item, Element, Unit, Flag) to keep the fact table narrow,
- merges value- and flag-streams so data-quality is always attached,
- fast-loads into the fact table. Using Foreach avoids five duplicate flows and ensures that adding a new continent is “drop file → run”.

4. Insights & Conclusions

a. Coverage of FAO reports has increased since 1961

Even without summing tonnages we can see the data set expand: the number of rows that FAO publishes for the element Production rises sharply over time.

```
SELECT d.YearLabel ,

       COUNT(*)        AS resultCount

FROM   dbo.FactProduction fp

JOIN   dbo.DimElement  e ON e.ElementKey =
fp.ElementKey

JOIN   dbo.DimDate      d ON d.DateKey  = fp.DateKey

WHERE  e.ElementName = N'Production'

GROUP  BY d.YearLabel

ORDER  BY d.YearLabel;
```

	YearLabel	resultCow
1	1961	5691
2	1962	5714
3	1963	5526
4	1964	5490
5	1965	5335
6	1966	5224
7	1967	5250
8	1968	5262
9	1969	5220
10	1970	5057
11	1971	4947
12	1972	5064
13	1973	4952
14	1974	5082
15	1975	5009
	...	
42	2002	6470
43	2003	6272
44	2004	6196
45	2005	6256
46	2006	6392
47	2007	6239
48	2008	6208
49	2009	6288
50	2010	6342
51	2011	6304
52	2012	6394
53	2013	6415
54	2014	6469
55	2015	6695
56	2016	6654
57	2017	6588
58	2018	6424
59	2019	7224

b. Asia dominates the modern dataset

Counting Production rows per continent for a recent year (2019) shows Asia produces, and therefore reports on, far more crop categories than any other region.

```
SELECT a.Continent ,

       COUNT(*) AS resultcount2019

FROM   dbo.FactProduction fp

JOIN   dbo.DimElement   e ON e.ElementKey =
fp.ElementKey

JOIN   dbo.DimArea      a ON a.AreaKey  = fp.AreaKey

JOIN   dbo.DimDate      d ON d.DateKey  = fp.DateKey

WHERE  e.ElementName = N'Production'

  AND  d.YearLabel   = N'2019'

GROUP  BY a.Continent

ORDER  BY resultcount2019 DESC;
```

	Continent	resultcount2019
1	Africa	2644
2	Asia	1805
3	Americas	1688
4	Europe	576
5	Oceania	511

c. Data quality is overwhelmingly “official”

The flag dimension attached during ETL lets us audit data reliability.

```
SELECT f.FlagCode ,

       f.FlagMeaning ,

       COUNT(*)        AS resultcount ,

       100.0*COUNT(*) / SUM(COUNT(*)) OVER() AS
PctRows

FROM   dbo.FactProduction fp

JOIN   dbo.DimFlag f ON f.FlagKey = fp.FlagKey

GROUP  BY f.FlagCode , f.FlagMeaning

ORDER  BY resultcount DESC;
```

	FlagCode	FlagMeaning	resultcount	PctRows
1	Fc	Calculated data	668662	48.002957730300
2	F	FAO estimate	273484	19.633298874339
3	A	Aggregate, may include official, semi-official, ...	185441	13.312729726625
4	Im	FAO data based on imputation methodology	134474	9.653830691477
5	M	Data not available	83949	6.026662646450
6	*	Unofficial figure	46950	3.370520330806

d. Conclusion

FAO's yearly coverage expands massively after 1960, Asia leads the growth, African yield reporting lags, and the vast majority of figures are classified "official".

The star-schema and ETL therefore deliver a clean, well-documented warehouse that supports rapid, trustworthy analysis—even when numeric aggregation is not possible—while making any future FAO file drop-in ready through the Foreach-Loop design.